STRUCTURAL BIOLOGY

ISSN: 2059-7983
journals.iucr.org/d

Probing the statistics of sequence-dependent DNA conformations in solution using SAXS

Heidar J. Koning, Anuradha Pullakhandam, Andrew E. Whitten, Charles S. Bond and Michel Peyrard

Acta Cryst. (2026). D**82**, 79–99

STRUCTURAL BIOLOGY

ISSN 2059-7983

Probing the statistics of sequence-dependent DNA conformations in solution using SAXS

Heidar J. Koning,[a] Anuradha Pullakhandam,[a] Andrew E. Whitten,[b] Charles S. Bond[a]* and Michel Peyrard[c]*

[a]School of Molecular Sciences, The University of Western Australia, 35 Stirling Highway, Crawley, WA 6009, Australia, [b]Australian Nuclear Science and Technology Organisation (ANSTO), New Illawarra Road, Lucas Heights, NSW 2234, Australia, and [c]Laboratoire de Physique CNRS UMR 5672, Ecole Normale Supérieure de Lyon, 46 Allée d'Italie, 69007 Lyon, France. *Correspondence e-mail: charles.bond@uwa.edu.au, michel.peyrard@ens-lyon.fr

Received 18 October 2024
Accepted 22 December 2025

Edited by S. Wakatsuki, Stanford University, USA

Keywords: DNA bending; polymer model; SAXS; AT tracts; GAGE6.

SASBDB references: GAGE6, SASDXH7; GAGE6_1, SASDV87; GAGE6_2, SASDV97; GAGE6_3, SASDVA7

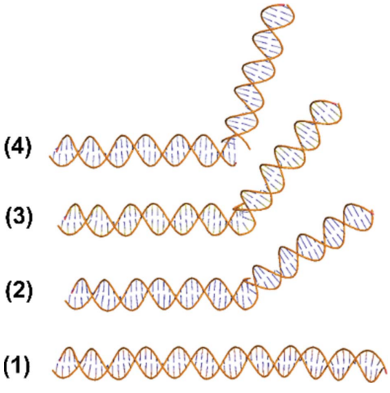
Supporting information: this article has supporting information at journals.iucr.org/d

SAXS studies of four 60-base-pair DNA duplexes with sequences closely related to part of the GAGE6 (G-antigen 6) promoter have been performed to study the role of DNA conformations in solution and their potential relationship to DNA–protein binding. We show that the SAXS data can be analysed using a simple polymer model, which nevertheless quantitatively describes the average persistence length and torsional rigidity of the DNA double helix, to determine the statistical distribution of local conformations of the DNA in solution to high accuracy. Although the SAXS data are averaged over time and all spatial orientations of the molecules, for sequences which have some asymmetry in the data we show that the conformations can be oriented with respect to the sequence. This allows specific features detected by the analysis to be precisely related to the DNA sequence, opening up new opportunities for SAXS to investigate the properties of DNA in solution. The biological implications of these results are discussed.

1. Introduction

In cells, DNA interacts with complex auxiliary machinery which allows its compact packing (Lowary & Widom, 1998), reading of the genetic code (Heumann *et al.*, 1988) or the control of gene expression (Knott *et al.*, 2016). Establishing extensive contacts between a protein and a rigid double helix is an improbable event. To mitigate this, these interactions commonly require the bending of DNA, as typified by DNA wrapping around nucleosomes (Lowary & Widom, 1998) but also when RNA polymerase binds to the double helix (Heumann *et al.*, 1988). To maximize the number of interactions between a protein and DNA, the DNA must be able to form surfaces complementary to binding regions or pockets in proteins. Thus, DNA curvature, flexibility and plasticity play an important role, particularly in the selection of a preferred protein binding site (Lowary & Widom, 1998; Anselmi *et al.*, 1999). This has led to many high-resolution structural investigations of free DNA and DNA–protein complexes (Crothers, 1988). Since the pioneering work of Dickerson & Drew (1981), which revealed the intrinsic curvature of a DNA dodecamer, many sequences with intrinsic curvature or twist have been found and tabulated (Olson *et al.*, 1998), leading to the idea of a 'spatial code hidden within the double helix' (Gorin *et al.*, 1995).

However, crystallographic and cryoEM studies have limitations because they do not reflect the dynamics of the DNA molecule in solution. The stiffness of DNA does not prevent it from showing large-scale fluctuations or local fluctuational openings which may strongly affect its local



flexibility (Theodorakopoulos & Peyrard, 2012). These structural dynamics are important both at the global and local levels and may provide molecular signals for protein binding which are not apparent in crystallographic and cryoEM data (Olson *et al.*, 1998). Molecular-dynamics (MD) simulations have been used to investigate the effects of the sequence of DNA on its flexibility. MD has allowed the detailed study of local distortions associated with bending at the scale of a few base pairs (Lankaš *et al.*, 2003) and may provide insight into the local and global effects of strong bending (Curuksu *et al.*, 2009). These prior studies establish a theoretical framework for understanding the conformational dynamics of dsDNA. However, validating these theoretical structural mechanisms requires that they explain a wide range of experimental observations from a wide range of structural and biophysical techniques.

Collecting and analysing data on DNA shape fluctuations in solution is not easy and is often performed via indirect approaches, such as gel migration or cyclization kinetics; techniques that lack the spatial resolution to study the effect of the sequence. Moreover, gels apply external constraints on the molecules and cyclization requires peculiar sequences or specific molecular constructs. Light-scattering studies provided the first measurements of the persistence length of DNA molecules that are unconstrained in solution (Peterlin, 1953). Such experiments used visible light and did not have sufficient spatial resolution to investigate the properties of DNA on the scale of a few base pairs. However, small-angle X-ray scattering (SAXS) offers a drastic increase in the resolution of experiments relative to light scattering. Despite this, extracting the high-resolution properties of DNA in solution from SAXS data is not straightforward, and this is one aspect that we discuss here.

A limitation of previous crystallographic and cryoEM studies focusing on DNA is that they only capture short-range effects. These studies often tabulate the properties of dimers of base pairs (Olson *et al.*, 1998), or even up to tetranucleotides (Lavery *et al.*, 2010), but, as we show in this work, the change in the mechanical response of a DNA molecule in which the sequence is locally modified may be more global and extend well beyond a few base pairs. Being able to detect the most likely conformations of a molecule of a few tens of base pairs in solution with a sufficient resolution is an important basis for the discussion of protein–DNA interactions with specific sequences.

In this work, we study a series of four DNA sequences closely related to the G-antigen 6 (GAGE6) oligonucleotide, which is a 60 bp dsDNA oligonucleotide that was first isolated by Song *et al.* (2005) as a fragment of the 2241 bp GAGE6 promoter. This 60 bp region was the only part of the digested promoter which bound the multi-functional nuclear protein Ptb-associated splicing factor (PSF), more commonly known as splicing factor proline/glutamine rich (SFPQ) (Song *et al.*, 2005). The DNA recognition mode of SFPQ is currently unknown, although the arginine–glycine (RGG)-rich region of the N-terminal intrinsically disordered region has been identified as the DNA-binding domain (Urban *et al.*, 2002; Lee *et al.*, 2015; Wang *et al.*, 2022). The stringent binding of SFPQ to the GAGE6 promoter perhaps suggests that elements of the DNA sequence create a specific binding site for SFPQ.

The native GAGE6 oligonucleotide sequence contains three AT-rich tracts of varying lengths. AT-rich tracts, in general, have been the subject of multiple studies focused on investigating whether they introduce additional flexibility to the duplex or are energetically capable of causing it to kink and bend in solution in the absence of protein binding (Chirico *et al.*, 2001; Haran & Mohanty, 2009; Harteis & Schneider, 2014; Hizver *et al.*, 2001; Schindler *et al.*, 2018; Stefl *et al.*, 2004; Widom, 1984). It is possible that AT-rich regions in the GAGE6 fragment trigger fluctuational opening of the base pairs or cause metastable kinks and bends to form in the duplex. These factors could in turn facilitate enhanced base access for SFPQ. Direct access to bases in both DNA or RNA is an important feature for RGG-containing binding domains and their interactions with nucleic acids (Chong *et al.*, 2018). To probe the effect of AT-rich sequences on the GAGE6 oligonucleotide, and to test the hypothesis that AT-rich domains create fluctuational opening of the base pairs and localized bending, we synthesized four 60 bp dsDNA oligonucleotides in which the AT tracts in the native GAGE6 sequence are incrementally swapped for GC-rich tracts (termed GAGE6, GAGE6_1, GAGE6_2 and GAGE6_3; Fig. 1). These were then studied using size-exclusion chromatography coupled with small-angle X-ray scattering (SEC-SAXS).

In this work, we show the following using the four GAGE6 variants.

(i) The local conformation (bending, twist) of a 60 bp DNA sequence in solution can be derived from the analysis of SAXS data using a simplified polymer model, which quantitatively

```
          .........1.........2.........3.........4.........5.........6
GAGE6   : 5-GCCTTCTGCAAAGAAGTCTTGCGCATCTTTTTGTGAAGTTTATTTCTAGCTTTTTGATGCT-3
GAGE6_1: 5-GCCTTCTGCAAAGAAGTCTTGCGCATCGCGCGTGAAGTTTATTTCTAGCTTTTTGATGCT-3
GAGE6_2: 5-GCCTTCTGCAAAGAAGTCTTGCGCATCGCGCGTGAAGGCGCGCGCTAGCTTTTTGATGCT-3
GAGE6_3: 5-GCCTTCTGCAAAGAAGTCTTGCGCATCGCGCGTGAAGGCGCGCGCTAGCGCGCGGATGCT-3
```

Figure 1
DNA sequences. The domains with at least four consecutive AT pairs in the GAGE6 sequence are highlighted in yellow. The grey rectangles mark those domains in which AT pairs have been replaced by GC pairs in the sequences GAGE6_1, GAGE6_2 and GAGE6_3. The top line marks ensembles of ten base pairs in the sequences.

Table 1
Sample information and characterization: small-angle X-ray scattering parameters.

	GAGE6	GAGE6_1	GAGE6_2	GAGE6_3
(*a*) Sample details				
Organism	*Homo sapiens*	Variant of original GAGE6 sequence	Variant of original GAGE6 sequence	Variant of original GAGE6 sequence
Source	Wang *et al.* (2022)			
Scattering-particle composition				
DNA/RNA(s)	60 bp dsDNA	60 bp dsDNA	60 bp dsDNA	60 bp dsDNA
Stoichiometry of components	Single component	Single component	Single component	Single component
Sample environment/configuration				
Solvent composition	150 m*M* KCl, 20 m*M* HEPES pH 7.4, 1 m*M* DTT, 5% glycerol, 5 m*M* $MgCl_2$	150 m*M* KCl, 20 m*M* HEPES pH 7.4, 1 m*M* DTT, 5% glycerol, 5 m*M* $MgCl_2$	150 m*M* KCl, 20 m*M* HEPES pH 7.4, 1 m*M* DTT, 5% glycerol, 5 m*M* $MgCl_2$	150 m*M* KCl, 20 m*M* HEPES pH 7.4, 1 m*M* DTT, 5% glycerol, 5 m*M* $MgCl_2$
Sample temperature (°C)	25	25	25	25
In-beam sample cell	Co-flow	Co-flow	Co-flow	Co-flow
Size-exclusion chromatograpy (SEC-SAXS)				
Sample-injection concentration (µ*M*)	69.7	70	75	81
Sample-injection volume (ml)	0.05	0.05	0.05	0.05
SEC column type	Superdex 200 5/150	Superdex 200 5/150	Superdex 200 5/150	Superdex 200 5/150
SEC flow rate (ml min^{-1})	0.4	0.4	0.4	0.4
(*b*) SAXS data collection				
Data-acquisition/reduction software	*scatterBrain* 2.82	*scatterBrain* 2.82	*scatterBrain* 2.82	*scatterBrain* 2.82
Source/instrument description or reference	Australian Synchrotron SAXS/WAXS beamline	Australian Synchrotron SAXS/WAXS beamline	Australian Synchrotron SAXS/WAXS beamline	Australian Synchrotron SAXS/WAXS beamline
Wavelength (nm)	0.10781	0.10781	0.10781	0.10781
Camera length (mm)	3000	2385	2385	2385
Measured q-range (q_{min}–q_{max}) ($Å^{-1}$)	0.0045–0.49	0.0068–0.62	0.0068–0.62	0.0068–0.62
Method for scaling intensities	Absolute scaling against water	Absolute scaling against water	Absolute scaling against water	Absolute scaling against water
Exposure time(s), No. of exposures	Frames 183–194 selected (12 × 1 s frames)	Frames 156–162 selected (7 × 1 s frames)	Frames 163–164 selected (2 × 1 s frames)	Frames 166–172 selected (7 × 1 s frames)
(*c*) SAXS-derived structural parameters				
Methods/software	*ATSAS* 4.0	*ATSAS* 4.0	*ATSAS* 4.0	*ATSAS* 4.0
Guinier analysis				
$I(0) \pm s$ (cm^{-1})	0.005 ± 0.00013	0.015 ± 0.00026	0.013 ± 0.00028	0.016 ± 0.00018
$R_g \pm s$ (Å)	49.89 ± 3.55	60.22 ± 2.23	54.21 ± 2.55	56.37 ± 1.52
min ≤ qR_g ≤ max limit (or data-point range)	0.37–0.96	0.52–1.01	0.42–1.06	0.38–1
Linear fit assessment (fidelity in *PRIMUS*)	1	0.71	0.9	0.97
Point range	5–22	5–22	5–22	5–22
PDDF/$P(r)$ analysis	*ATSAS* 3.2.1	*ATSAS* 3.2.1	*ATSAS* 3.2.1	*ATSAS* 3.2.1
$I(0) \pm s$ (cm^{-1})	0.00509 ± 0.0001516	0.01459 ± 0.0001741	0.01250 ± 0.0002808	0.01554 ± 0.0002117
$R_g \pm s$ (Å)	55.18 ± 1.829	54.24 ± 1.031	55.37 ± 1.860	55.90 ± 1.380
d_{max} (Å)	195	192	184.7	215
q-range ($Å^{-1}$)	0.0080–0.1704	0.0104–0.1554	0.0077–0.1518	0.0068–0.1536
$P(r)$ fit assessment (total quality estimate)	0.67 (reasonable)	0.70 (reasonable)	0.67 (reasonable)	0.68 (reasonable)
α	1.308	0.552	0.413	1.48
(*d*) Scattering-particle size				
Methods/software	*ATSAS* 3.2.1	*ATSAS* 3.2.1	*ATSAS* 3.2.1	*ATSAS* 3.2.1
Volume estimates				
Volume ($Å^{-3}$) (MoW method)	43073	69629	69523	79329
Molecular-weight estimates (kDa)				
From chemical composition	37.101	37.105	37.111	37.116
From SAXS concentration-independent method (volume of correlation method)	30.06	31.61	35.53	34.95
(*e*) Data deposition				
SASBDB code	SASDXH7	SASDV87	SASDV97	SASDVA7

describes both DNA persistence length and torsional rigidity. This model allows a very broad exploration of the conformational space of the sample and is able to detect subtle effects with a resolution on the scale of a few base pairs.

(ii) The effect of the sequence on the conformation of dsDNA in solution is nontrivial and may not reduce to a sum of local effects. Changing a flexible polynucleotide segment into a more rigid one may, for instance, strongly enhance the bending fluctuations in another domain.

2. Materials and methods

2.1. Synthesis of oligonucleotides

The 60 bp HPLC-purified dsDNA oligonucleotides GAGE6, GAGE6_1, GAGE6_2 and GAGE6_3 (Fig. 1) were produced by Integrated DNA Technologies and resuspended in 150 m*M* KCl, 20 m*M* HEPES pH 7.4, 1 m*M* DTT, 5%(*v*/*v*) glycerol, 5 m*M* $MgCl_2$ buffer to a DNA concentration of 70–80 µ*M* (exact concentrations are specified in Table 1).

2.2. Size-exclusion chromatography–small-angle X-ray scattering (SEC-SAXS)

SAXS data for all oligonucleotides were collected on the SAXS/WAXS beamline at the Australian Synchrotron using an inline SEC-SAXS sheath-flow setup (Kirby *et al.*, 2016; Ryan *et al.*, 2018). All data were collected by loading 50 µl of DNA at 70–80 µ*M* in a buffer consisting of 150 m*M* KCl, 20 m*M* HEPES pH 7.4, 1 m*M* DTT, 5%(*v*/*v*) glycerol, 5 m*M* $MgCl_2$. All samples were analysed on a pre-equilibrated Superdex 200 Increase 5/150 column (GE Healthcare) with UV absorbance at 260 and 280 nm monitored alongside scattering. Data reduction to $I(q)$ versus q scattering profiles was carried out using *scatterBrain* (software for acquiring, processing and viewing SAXS/WAXS data at the Australian Synchrotron), correcting for sample transmission and solvent scattering, and data were placed on an absolute intensity scale using a water standard. Subsequent data processing and analysis were performed using the *ATSAS* suite (Petoukhov *et al.*, 2012). As discussed by Trewhella *et al.* (2017), *scatterBrain* outputs the uncertainty of intensity measurements as 2σ. For subsequent analysis, these uncertainties were transformed to σ for all data sets such that all metrics used for comparing models and experimental data had conventional interpretations. Supplementary Fig. S4 shows that this does not affect the shape of the distance distribution functions $P_{exp}(r)$ which this method uses for analysis (see Sections 2.3 and 2.4). For all SEC-SAXS data, regions with self-consistent, non-nucleic acid frames prior to the elution of the main peak were averaged and taken as solvent scattering with the program *CHROMIXS* (Panjkovich & Svergun, 2018). The sample scattering was then taken as the average of frames with similar values of the radius of gyration R_g within the peak corresponding to DNA (Supplementary Fig. S1). To additionally assess the monodispersity of the frames across each chromatography peak, the molecular weight of the individual frames was calculated in *BioXTAS Raw* (Hopkins *et al.*, 2017). This was performed using the V_c method (Rambo & Tainer, 2013) with the molecule type set to 'RNA' rather than protein and the average window size set to 1 (Supplementary Fig. S2). Guinier analysis was performed using *ATSAS* 4.0 (Manalastas-Cantos *et al.*, 2021). Distance distribution analysis was performed using *ATSAS* 3.2.1 (Manalastas-Cantos *et al.*, 2021). To assist in the selection of the optimization parameters for our various $P(r)$ functions, the program *AutoGNOM*, which is part of the *ATSAS* package for the analysis of SAXS data (Petoukhov *et al.*, 2012), was also used. To assess the role of regularization in the stability of our main features in $P_{exp}(r)$, we performed a cross-comparison between the optimal solutions derived from *AutoGNOM*-4 from *ATSAS* 2.7.2 and *AutoGNOM*-5 from *ATSAS* 3.2.1.

2.3. Choosing real space instead of reciprocal space for analysis

Scattering experiments measure the intensity as a function of the scattering vector **q**. Coherent elastic scattering is determined by the correlations in the local scattering density of the sample and is expressed by the Fourier transform of the autocorrelation function of the scattering density. Moreover, for dilute molecular solutions, where the orientation and position of individual molecules are uncorrelated, the only coherent contribution to the scattering comes from the internal structure of the molecules. As the scattering represents the ensemble- and time-averaged scattering of all molecules illuminated by the beam, all orientational information is lost and the scattering patterns are radially symmetric. Thus, scattering data from dilute monodisperse solutions depends only on the modulus q of the scattering vector. The Fourier transform in polar coordinates reduces to a one-dimensional integral (Sivia, 2011)

$$I(q) \propto \int_0^{D_{max}} P(r) \frac{\sin qr}{qr} \, dr, \tag{1}$$

where $P(r)$ is the pair-distribution function of the scattering particle. This function can be related to the probability of finding two scattering points separated by a distance between r and $r + dr$. The shape of the $P(r)$ function is characteristic of the shape of the molecule and depends on the input parameter D_{max}, which is the maximum distance between two scattering points within a given molecule. Whilst D_{max} is selected and optimized during the fitting process, it also reflects a real physical property, which is the maximum linear dimension of the molecule. Data are collected in reciprocal space q, while we are interested in conformations, *i.e.* properties related to the geometry of the molecules in real space r. The pair-distance distribution function $P(r)$ can be derived from $I(q)$ by inverse Fourier transform,

$$P(r) \propto r^2 \int_0^{\infty} q^2 I(q) \frac{\sin qr}{qr} \, dq, \tag{2}$$

meaning that the data in real space r and reciprocal space q are related by a mathematical transform and are therefore formally equivalent. It is evident from equation (1) that knowledge of the $P(r)$ function permits the calculation of $I(q)$; however, from equation (2) it is clear that the inverse is not true. Computation of $P(r)$ from $I(q)$ involves an infinite integral, yet the scattering data are measured over a finite q-range; thus, a direct calculation of inverse Fourier transformation of $I(q)$ will be dependent on the q-range over which the data are measured. To address this issue, an indirect Fourier transformation of the scattering data is used, whereby a set of basis functions are combined to generate a $P(r)$ function, from which $I(q)$ can be calculated and compared with the experimental scattering data. The coefficients of the basis functions are then optimized (subject to various restraints and constraints) against the scattering data to yield the $P(r)$ function for the measured $I(q)$ data. The real-space representation of the scattering data is more intuitive to most people and, importantly, subtle differences in reciprocal-space $I(q)$ can often be more obvious in real space. To this point, Fig. 2 plots the theoretical $P(r)$ and $I(q)$ profiles of a short segment of DNA made of two straight double helices

connected by a variety of local bends. In the theoretical data the structural differences due to bending lead (Fig. 2*a*) to an accumulation of differences at specific r values and a notable qualitative change in the shape of $P(r)$ (Fig. 2*b*). Whilst the theoretical scattering for these different bent structures is only subtly different below $q = 0.23$ Å^{-1} (Figs. 2*c* and 2*d*), the scattering begins to show discernible differences above this threshold in the theoretical profiles (Fig. 2*d*). However, when considering real experimental data (see Fig. 5 for an example) above this threshold, the differences may become increasingly difficult to discern due to rising noise levels as the data extend along q. In contrast to the smooth theoretical curves, the experimental scattering profile also becomes much more diffuse at higher q, possibly reducing sensitivity to structural differences among such conformations when relying on this part of the scattering data. This example illustrates a crucial choice for the successful analysis of our SAXS data: to detect the structural properties of the DNA molecules, it is much more efficient and meaningful to fit $P(r)$ in place of $I(q)$. This can be easily understood from the general relationship between real space and reciprocal space. In this study, we are looking for DNA features which could affect the binding of proteins, such as curvature or flexibility on the scale of the binding domain. These features are *localized* structures in real space, corresponding to *extended* structures in the Fourier-transformed reciprocal space, where they are spread over a broad q-range, and therefore are harder to detect.

The radius of gyration can be calculated directly from $P(r)$ and can be generally determined more accurately when compared with the Guinier approximation (Glatter, 1977) as a significantly larger region of reciprocal space is used to calculate $P(r)$ (Blanchet & Svergun, 2013).

To perform a comparison of the measured data sets in reciprocal space the data were re-gridded in *PRIMUS* with a spacing of $\Delta q = 0.001$ Å^{-1}. The re-gridding was carried out using a common q-range between all four experiments of 0.0068–0.4918 Å^{-1}. The 'data compare' function in *PRIMUS* was then used to assess the statistical similarity of the data sets in reciprocal space. To focus on the q-range that gives rise to the features that we have observed in $P_{\rm exp}(r)$, the maximum value of q of the data used for the test was set to 0.18 Å^{-1}. It computes a p-value by checking for patterns in the cross-correlation matrix between data sets, which does not depend on the values of the experimental errors (Franke *et al.*, 2015). While $p > 0.95$ means that two data sets are almost identical, low values of p, typically below a threshold of 0.01–0.05,

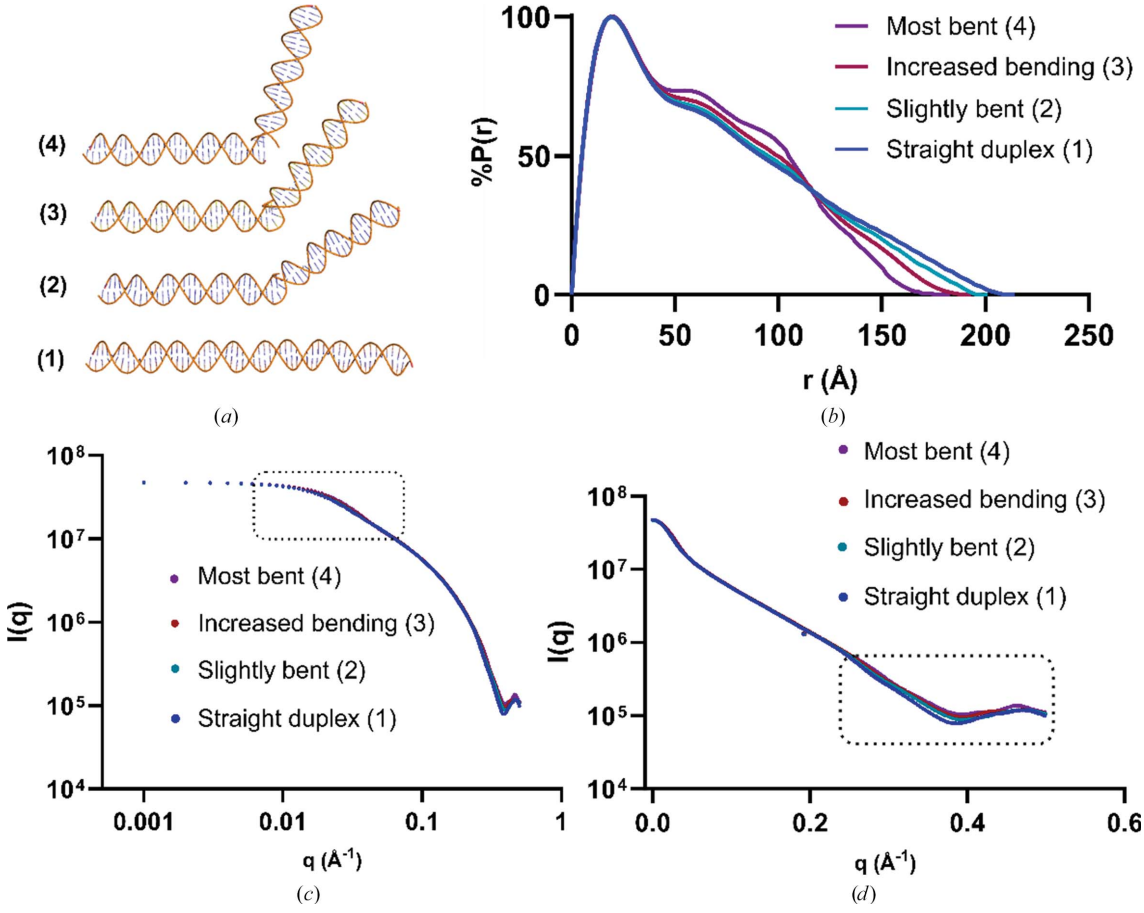


Figure 2
Illustration of the effect of DNA bending on the shape of $P(r)$. (*a*) A group of simple 60 bp DNA models, the first of which was created by *AlphaFold*3 (Abramson *et al.*, 2024). Models 2–4 were created by variably bending the first *AlphaFold* model manually at base pair 34. (*b*) Pair-distance probability distribution functions $P(r)$ for these four conformations computed with *CRYSOL* (Svergun *et al.*, 1995). (*c*, *d*) The simulated scattering from the four duplexes is shown as log(I) versus log(q) (*c*) or log(I) versus q (*d*). The areas where the computed scattering patterns differ across both types of plots are indicated by the dotted box.

indicate that the data sets are statistically independent. Such values are generally used to identify outliers, such as sample misloading, denaturing or radiation damage (Manalastas-Cantos *et al.*, 2021). Intermediate values of p indicate some correlations between the data sets. Supplementary Table S1 shows that the p-values obtained by comparing the GAGE6 data successively with data recorded for GAGE6_1, GAGE6_2 and GAGE6_3 are 0.493, 0.280 and 0.078, respectively. This indicates decreasing correlations between the data sets, which are consistent with the nature of our samples. They derive from the same GAGE6 sequence, with only local sequence changes, increasing from GAGE6_1 to GAGE6_3. Therefore, one should not expect that such samples would lead to completely statistically independent data sets as occurs for outliers due to experimental problems. These p-values are interesting because our analysis (see Section 3) detects a large bending domain in GAGE6_1 and GAGE6_2 in a portion of the sequence which is not far from the bent domain detected for GAGE6, which is compatible with a fairly large p-value, whereas this strongly bent domain disappears for GAGE6_3, which should decrease the correlation with the GAGE6 data set, leading to a lower p-value. The statistical analysis therefore appears to provide a very good preview of the results presented in Section 3.

2.4. Derivation of $P_{exp}(r)$ from experimental data

The derivation of the pair-distance distribution function (Supplementary Fig. S3) $P_{exp}(r)$ from $I(q)$ versus q via an indirect Fourier transform is performed with the program *GNOM* from the *ATSAS* package (Manalastas-Cantos *et al.*, 2021). The program requires several input parameters which can variably affect the results. Some of these relate to the experimental setup, and others have to be selected. Two of these selected parameters are crucial to the transform process: the maximum diameter of the sample D_{max} and the regularization parameter α.

The parameter D_{max} can be estimated to good accuracy from the properties of the samples. According to the typical base-pair distance of 3.3–3.4 Å in the B-form of DNA (Saenger, 1984), a fully rigid ideal 60 bp duplex should have a maximum dimension D_{max} of ∼205 Å (excluding its hydration shell). Using this as an estimate for what our approximate maximal dimension should be in solution, one of the criteria was that the selection of D_{max} for each function falls within 10% (184.5–225.5 Å) of this value, as there can often be an inherent uncertainty associated with D_{max} which can be difficult to quantify (Trewhella *et al.*, 2017). As shown by Supplementary Fig. S5, varying D_{max} over a range of 20 Å (which is very significant in comparison with the length of the DNA sequences) only has a minute effect on the shape of $P_{exp}(r)$. Furthermore, as small errors in D_{max} can alter $P(r)$ slightly (Grant *et al.*, 2015), persistence of features in the respective functions in the face of varying D_{max} serves as an additional indicator of the robustness of the solution.

The choice of the regularization parameter α is more delicate. The program *GNOM* includes a process called *Auto-GNOM* which optimizes a Total Quality Estimate (TQE) that combines an ensemble of different criteria (Svergun, 1992) such as the discrepancy between the reconstructed Fourier space data and the experimental data, the degree of oscillations in the calculated $P_{exp}(r)$ and the stability of the solution versus variation of α. A convergence process optimizes the TQE. Throughout the development of the *ATSAS* package the method has evolved, and currently the more contemporary version *GNOM*-5 has introduced a new smoothness criterion for $P_{exp}(r)$ that was not part of *GNOM*-4 but is now used in addition to the other metrics to calculate the TQE. The best definition of the TQE and the best convergence procedure may depend on the type of sample. It is likely that a sample with a well defined shape such as a folded protein may have different requirements from a long flexible polymer such as DNA, as these often are molecules that may have very different conformations in solution. Thus, in spite of the long development period of *ATSAS*, selection of the optimal α parameter remains a challenge. This is why we have compared the results provided by both *GNOM*-5 and *GNOM*-4. They are plotted in Supplementary Fig. S6, which also shows how the TQE of *GNOM*-4 depends on α for all of the samples.

For GAGE6 the results of *GNOM*-5 and *GNOM*-4 are very close to each other for the shape of $P_{exp}(r)$ and its estimated error bars. For GAGE6_1 and GAGE6_3 both optimization procedures converge towards similar shapes, although some differences are noticeable. For GAGE6_2, *GNOM*-5 converges to α = 0.01084, *i.e.* a very weak regularization, leading to $P_{exp}(r)$ showing significant oscillations and large error bars. *GNOM*-4 converges to α = 1.31, *i.e.* a fairly strong regularization with a very smooth $P_{exp}(r)$ and small error bars. Supplementary Fig. S6 shows that for GAGE6_2 the TQE has a wide plateau where it stays almost constant when α varies. This explains why determination of the optimal α is particularly difficult for this sample. Therefore, for some samples the calculation of $P_{exp}(r)$ raises some doubts concerning the validity of the results. However, Supplementary Figs. S6 and S7 show that α essentially determines the *magnitude* of the humps of $P_{exp}(r)$ but has little effect on their *position* with respect to r. This is important because Fig. 2 shows that the amplitude of a local DNA bend affects the magnitude of the humps in $P_{exp}(r)$ but not their position in r, which is determined by the location of the bend along the sequence.

This suggests that a possible indetermination of α could have little influence on the main DNA feature that we are trying to detect in our analysis, *i.e.* how the sequence of DNA locally modifies its propensity for bending, which may affect protein binding. As shown later, even for the case of sequence GAGE6_2, this conjecture is confirmed by the calculation presented in Section 3.3, although it is the sequence for which the evaluation of α and the shape of $P_{exp}(r)$ is the hardest. Supplementary Fig. S10 shows that the same is true for all sequences.

This shows that SAXS data for flexible polymers, which can hardly be distinguished from each other in reciprocal space, even with mathematical tools, can contain detailed information which is revealed when the results are expressed in real

space. The above factors, together with the simulated $P(r)$ functions, which show that bending a dsDNA model will cause bumps to appear in the function, the fact that AT tracts have been proven to cause bending of dsDNA, the good fits of each $P(r)$ function to the data (Supplementary Fig. S8) and their reasonable TQE scores, strongly support the validity of the information collected in real space. This is further confirmed by the correlation between the sequence of our DNA samples and the local properties that we detect in our analysis based on $P_{exp}(r)$.

Simulated $P(r)$ functions (Fig. 2) were generated using *CRYSOL* (Svergun *et al.*, 1995) on the first *AlphaFold*3 model and the three subsequent manually bent models, applying fake 1% errors to the theoretical intensity and proceeding with the standard derivation of $P(r)$ in *PRIMUS*.

2.5. Analysis: from $P_{exp}(r)$ to the conformation of the samples

2.5.1. General concept

The pair distribution function, which describes the probabilities of all of the atom–atom distances in the conformational ensemble of a sample, summarizes a large amount of information. Even in the absence of conformational heterogeneity, *ab initio* structural modelling against high-quality data from monodisperse solutions does not yield unambiguous structures. Add the complication of conformational heterogeneity, and the extraction of structural and conformational information from scattering data is not possible in the absence of complementary information from other biophysical techniques. Indeed, there are numerous packages available for modelling ensembles of flexible molecules using SAXS data, but all rely on having high-resolution structural information for all components in the flexible system (Bernadó *et al.*, 2007; Schneidman-Duhovny *et al.*, 2016; Tria *et al.*, 2015). Similarly to flexible protein solutions (Martin *et al.*, 2021), DNA in solution is an ensemble of different structures that are evolving over time, and beyond a few helical turns a DNA molecule undergoes large-scale fluctuations. Our analysis proceeds along similar lines, taking advantage of the knowledge of the general properties of DNA to build a suitable ensemble of conformers. Our method was first developed to search for kinked DNA conformations in SAXS and SANS data on a nucleosome-positioning sequence (Schindler *et al.*, 2018). However, it has been significantly improved for the present study with the set of data obtained on a series of related sequences, with additional steps now added to the method. A polymer model of DNA is used to widely scan the conformational space of the sample. Ensembles of conformations which have a pair-distance distribution probability sufficiently close to $P_{exp}(r)$ are selected and saved. A subsequent analysis then selects a subset of optimal conformations, which are then statistically analysed to extract the main features of the DNA sample such as its local distribution of bending or twist angles and their standard deviations. In some cases, small differences in the sequence are found to lead to large changes in sample properties, which might be relevant for protein binding affinity or the selection of a binding site. Although a random scan of conformations is, in principle, possible provided that it is sufficiently broad, it would be highly inefficient. The choice of a suitable polymer model that reflects the properties of DNA is important. However the model must not bias the results, and therefore it does not include any assumption on the local properties of the sample.

The approach used by Peterlin (1953) to extract the persistence length of DNA from light-scattering data gives some hints on possible approaches for SAXS data. Peterlin introduced a simple polymer model, the Kratky–Porod model, which consists of a chain of rigid segments connected by flexible joints with an energy that depends on the angle θ between the two joined segments. For this simple model, theoretical calculation of the scattering function is possible within some limits and its comparison with the experimental data determines the appropriate model parameter, from which the persistence length can be derived. The same idea was refined by Schellman (1974) for a better description of DNA flexibility which took into account dihedral angles φ between the planes defined by pairs of consecutive segments. However, to allow analytical calculations, he only considered two limiting cases: either φ is constrained to zero, *i.e.* the model evolves in a plane, or φ is totally random. For DNA the reality lies between these two limits, which correspond to either an infinite torsional rigidity ($\varphi = 0$) or zero torsional rigidity (φ random, evolving freely without any energy cost). This is why we have chosen an extended Kratky–Porod model with an added energy contribution for the dihedral angles in order to describe not only the persistence length of DNA but also its proper torsional rigidity.

2.5.2. The DNA polymer model

The DNA polymer model is shown in Fig. 3. It is not concerned with the internal structure of the DNA helix but only with the conformation of the line going through the centre of the double helix. It consists of N objects representing the base pairs, each separated by the base-pair distance in DNA, a = 3.34 Å. Its $N - 1$ segments can be viewed as a generalized Kratky–Porod model. Its conformation is defined by the bending angles θ_n, $n = 2, \ldots, N - 1$ at sites n, between segment $n - 1$ and n ($-\pi < \theta_n \leq \pi$), and the dihedral angles

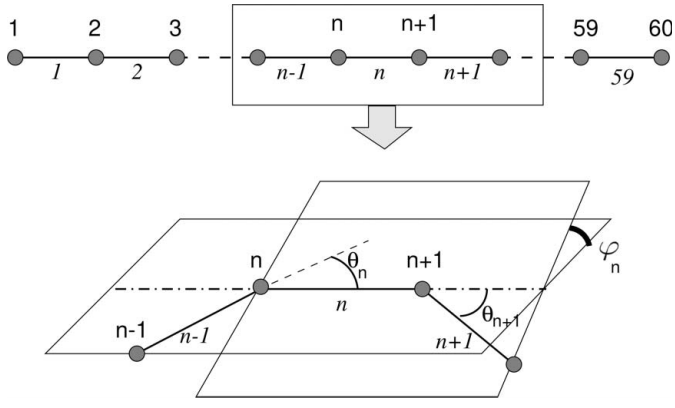


Figure 3
The DNA polymer model used in the analysis of the SAXS data.

φ_n, $n = 2, \ldots, N-2$, between the planes defined by segments $[n-1, n]$ and segments $[n, n+1]$ ($-\pi/2 < \varphi_n \leq \pi/2$). The reference state for this model is a straight DNA helix, with 11 base pairs per turn, *i.e.* an angle of about 33° between them. The θ and φ angles describe the deviations with respect to this structure.

The thermal fluctuations of the backbone are controlled by the variation of its bending and torsional energies E_θ and E_φ,

$$H = E_\theta + E_\varphi = \sum_{n=2}^{N-1} K(1-\cos\theta_n) + \sum_{n=2}^{N-2} C(1-\cos\varphi_n). \quad (3)$$

The constants K and C set the scale of the bending and torsional energies. They are assumed to be the same along the full model, and the bending and twist energies are minimal for $\theta = 0$ and $\varphi = 0$, *i.e.* the model is homogeneous, with no permanent bending or torsion, in order to avoid any bias towards a specific conformation other than a straight, untwisted, DNA molecule.

The numerical calculations use dimensionless variables. Lengths are measured in units of $a = 3.34$ Å, the DNA base-pair distance, so that the dimensionless length of the segments is unity. The energy is measured in units of $k_B T_{room}$, where k_B is the Boltzmann constant and T_{room} is room temperature in kelvin, chosen as 298 K. The two model parameters K and C are chosen to correspond to the physical properties of DNA.

The constant K determines the persistence length of the model, *i.e.* the length L_p over which unit vectors $\mathbf{R}_1$, $\mathbf{R}_2$ tangent to the helix axis lose collinearity according to $\langle \mathbf{R}_1 \cdot \mathbf{R}_2 \rangle = \langle \cos\theta \rangle \simeq \exp(-l_{12}/L_p)$, where $\langle\rangle$ designates a statistical average and l_{12} is the distance along the molecule between the two contact points of $\mathbf{R}_1$ and $\mathbf{R}_2$. It has been determined by a broad variety of methods, from macroscopic evaluations (Peterlin, 1953) to single-molecule experiments (Smith *et al.*, 1992) and, although it depends on external conditions (Lu *et al.*, 2002), the consensus for standard conditions is $L_p = 500$ Å, *i.e.* in the unit length of our model $L_p = 500/3.34 \simeq 150$ units. For the Kratky–Porod model, in the limit of L_p much larger than the length of a segment, L_p is given by $L_p = aK/k_B T$, so that, with our dimensionless variables where a and $k_B T$ are unity (for experiments at room temperature), we obtain $L_p = K$. Therefore, the choice $K = 150$ gives a model with a persistence length appropriate for DNA.

The constant C determines the torsional rigidity ξ of the model in the linear limit, *i.e.* for small torsional angles φ for which $E_\varphi \simeq (1/2)C\varphi^2$ and a torque $\tau = \partial E_\varphi/\partial\varphi = C\varphi$. In this limit a DNA molecule of length L behaves as an elastic rod whose torsional torque grows with torsion as $\tau = (\xi/L)\varphi$. For DNA, measurements of ξ vary from about 200 to 480 pN nm^2. Topoisomer distributions of small circles, yielding $\xi \simeq 300$ pN nm^2, are generally considered as the most reliable measurements (Bryant *et al.*, 2003). In the context of our model, for a base pair, *i.e.* $L = a = 3.34$ Å, we obtain $C = \xi/a = 898$ pN nm. When this value, homogeneous to energy (force × length), is expressed in our energy unit $k_B T_{room}$ we obtain the dimensionless value $C \simeq 218$. Taking into account the broad error bars in the experiments, we have used the value $C = 200$ in our calculations as a realistic value to describe the torsional modulus of DNA.

Conformations of this model at room temperature are generated using Monte Carlo simulations. The first segment $\mathbf{R}_1$ is chosen arbitrarily along the z axis of our reference frame, and a second segment $\mathbf{R}_2$ is added in the x, z plane by selecting a random value of θ_2 in the range $-\pi < \theta_2 \leq \pi$. From two vectors we can generate a local frame x', y', z', using $\mathbf{R}_2$ as the z' axis, choosing an orthogonal x' axis in the $\mathbf{R}_1, \mathbf{R}_2$ plane and completing by a third axis y' to obtain an orthonormal frame. A new segment is added in this local frame by selecting θ in the range $-\pi < \theta \leq \pi$ and the dihedral angle φ in the range $-\pi/2 < \varphi \leq \pi/2$, and so on until the 59 segments corresponding to our 60-base-pair DNA samples have been added. Each time a new segment has been defined the contribution $E_\theta + E_\varphi$ that it adds to the energy is compared with $k_B T$ ($k_B T = 1$ in our dimensionless units) and accepted or rejected with the Metropolis algorithm (Metropolis *et al.*, 1953). Once a conformation of 59 segments has been completed, the probability distribution $P(r)$ of the pair distances between its 60 beads is obtained by a calculation which directly derives from its definition. We compute all pair distances r_{ij}, $i = 1, \ldots, N$, $j = 1, \ldots, N$, $i \neq j$, *i.e.* $N(N-1)/2 = 1770$ values. Then, $P(r)$ is derived from a normalized histogram of these distances by counting how many of them lie within each of $N+1$ boxes of size a centred on the values $r_k = (k - \frac{1}{2})a$ (we use real distances, not the dimensionless distances for comparison with experimental data). $P(r_k) = n_k / \sum n_k$, where n_k is the number of r_{ij} values that fall within box k. The least-squares distance S between $P(r)$ and $P_{exp}(r)$, the normalized distribution function obtained from experimental data, interpolated to obtain its values at points r_k, is given by $S = \{[\sum_k [P(r_k) - P_{exp}(r_k)]^2\}^{1/2}$.

In this sum, only values $r_k > 25$ Å are considered to take into account the limitations of the DNA model, restricted to the backbone, which cannot describe DNA at very small distances. The experimental pair-distance probability distribution function $P_{exp}(r)$ contains many short-distance contributions which are not relevant for the overall conformation of the molecule. These include the distances between atoms belonging to the same base pair for which $r \leq d \simeq 25$ Å, where d is the diameter of the molecule, including bound water. Other short distances bind atoms belonging to neighbouring or next-neighbouring base pairs, which are also within $r \lesssim 25$ Å. The model, which represents one base pair by a single point, does not include all of these contributions, and even slightly above this threshold it tends to underestimate $P_{exp}(r)$. On the other hand, for larger distances the backbone gives a meaningful picture of the conformation of a DNA molecule and its flexibility, which is the aim of our analysis. If S is smaller than a predefined threshold S_0 the configuration is saved in a file for further analysis. To make sure that we only keep the conformations that provide an optimal fit to our experimental data, S_0 is selected so that only 10–50 conformations are saved for one million generated conformers. To ensure the thorough exploration of the conformational space of samples a calculation can generate up to 5×10^8 trial conformations, which is possible even with modest computing

facilities because the model is sufficiently simple, with only two parameters per DNA base pair. The same analysis with an all-atom DNA model would be beyond even the best computing facilities.

2.5.3. Analysis of the saved conformations

Among the saved conformations, we select the 1000 which have the smallest least-square difference S between $P(r)$ and $P_{\rm exp}(r)$, and we use these 'best' conformations for statistical analysis. An individual conformation has little physical meaning because at room temperature a 60 bp DNA molecule is sufficiently flexible to allow its shape to fluctuate widely. Nevertheless, its properties are reflected in the statistics of these fluctuations.

In the context of protein binding, DNA curvature and flexibility are important, therefore the first relevant set of statistics concern the bending angles $\theta_n(i)$, where n is the index of base pairs along the sequence and i (i = 1, ..., 1000) denotes the conformations.

To demonstrate these principles we foreshadow in Fig. 4 an example of the average over the 1000 best conformations of the absolute values $|\theta_n(i)|$ of the local bending angles, which measure the deviations from a straight double helix, *i.e.*

$$\overline{|\theta_n(i)|} = \frac{1}{1000}\sum_{i=1}^{1000} |\theta_n(i)|, \tag{4}$$

as well as their standard deviations

$$\sigma_\theta = \left[\frac{1}{1000}\sum_{i=1}^{1000}\left(|\theta_n(i)| - \overline{|\theta_n(i)|}\right)^2\right]^{1/2}. \tag{5}$$

We also show the average of $|\theta_n(i)|$ restricted to bending angles which exceed a threshold $\theta_0 = 5^\circ$ to stress the domains which are prone to large bending. The average dihedral angle $\overline{|\varphi_n(i)|}$ and the standard deviation σ_φ are computed with equations similar to equations (4) and (5).

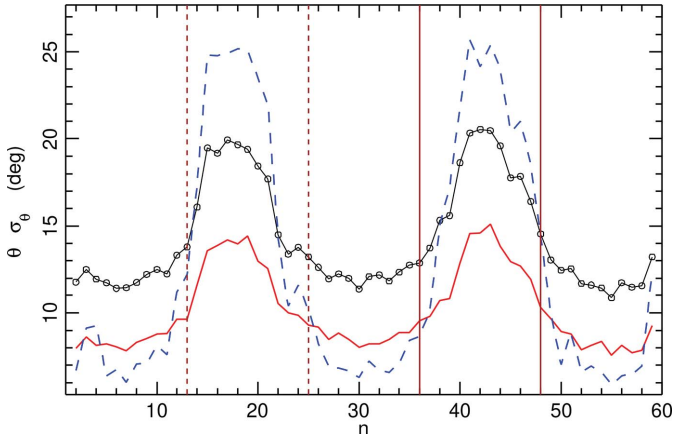


Figure 4
Statistics of the local bending angles $|\theta_n(i)|$ over the 1000 best saved conformations for the GAGE6 sample: the red full line shows $\overline{|\theta_n(i)|}$. The black line with circles shows the same average limited to the bending angles which are above $\theta_0 = 5$, and the dashed blue curve shows the standard deviation of $|\theta_n(i)|$. The full brown vertical lines show the positions $n_1 = 36$, $n_2 = 48$ which have been used for the 'orientation' of the conformations and the dashed brown vertical lines show the mirror positions $n_1' = 13$ and $n_2' = 25$.

Fig. 4 shows a typical pattern from SAXS results: it is symmetrical with respect to the centre of the sequence. This was expected because the structure factor and the probability distribution function $P_{\rm exp}(r)$ result from an average of all of the possible orientations of the molecules in space. Analysing $P_{\rm exp}(r)$ in terms of conformations has already reduced this three-dimensional degeneracy and brought the data along a line corresponding to the DNA chain. However, SAXS data alone do not allow us to identify a particular end of the sequence. To discuss the results, we would like to compare the sequence with its base pairs identified by an index p, p = 1, ..., 60 and the angles along the model $|\theta_n|$, n = 1, ..., 60, but we face a dilemma that the experiment cannot solve: it cannot tell us whether $p = 1$ corresponds to $n = 1$ or whether it corresponds to $n = 60$. The two peaks may be real, or one of them may be an artefact because the statistical analysis mixes sequences with $n = p$ and others with $n = 60 - p + 1$ so that we sum up data with a single peak on the right and data with a single peak on the left.

However, if the sequence has some intrinsic asymmetry this provides an additional piece of information which can be combined with the SAXS data to solve the dilemma by examining the conformations one by one and reversing some of them so that the peak of large $|\theta_n|$ is always on the same side, for instance to the right, which we call the 'forward' conformation. The algorithm to achieve what we call the 'orientation' of the conformations is the following. We identify the indices n_1, n_2 on both sides of the large n peak, for instance 36 and 48 in the case of the GAGE6 sample in Fig. 4. The mirror positions are $n_1' = N - n_2 + 1$ and $n_2' = N - n_1 + 1$, *i.e.* $n_1' = 13$ and $n_2' = 25$ for the example above. We then scan the 1000 best conformations again and compute for each of them $s = \sum|\theta_n|, n = n_1 \ldots n_2$ and $s' = \sum|\theta_n|, n = n_1' \ldots n_2'$. If the difference between s and s' is significant, which is evaluated by calculating $\Delta s/s = |s' - s|/s$, then one of the two peaks is ficticious. If $s > s'$ this conformation is already in the forward orientation. If $s' > s$ the conformation has to be reversed by switching indices n into $n' = N - n + 1$. After this check, constructing a plot similar to Fig. 4 with all conformations oriented 'forward', only one of the two peaks remains. As shown in Section 3, this is what happens for the four samples that we studied. When reversing a conformation, we also reverse the indices for the dihedral angles φ_n, paying attention to the fact that while n varies from 2 to $N - 1$ for θ_n, it varies from 2 to $N - 2$ for φ_n, so that instead of n' one has to use $n'' = N - n$ for φ. Besides the elimination of spurious signals, the orientation process improves the quality of the averaging because it becomes real averaging over the 1000 best configurations instead of the superposition of two sets of data averaged over about 500 conformations.

After this process has been completed, we know that all of the conformations generated by the model have the same orientation with respect to the sequence, *i.e.* that for all of them $n = 1$ corresponds to $p = 1$ or all of them correspond to $p = 60$. It does not yet tell us which of the two choices is

correct, and further data processing cannot answer that. However, now that we have a reliable statistical analysis we can compare $\overline{|\theta_n(i)|}$ with the sequence. We can expect that the domains with large bending angles are more likely to be found in regions which are richer in AT base pairs and are usually more flexible than GC-rich regions. For instance, for the sequence GAGE6 (Section 3.2), after orientation we detect a single peak around sites 38–44 (Fig. 6). In the sequence the domain p = 38–44 is a large continuous domain with only AT pairs. This suggests that for this sequence the correct choice is $n = p$, *i.e.* the case for which $n = 1$ corresponds to $p = 1$.

This 'orientation' step is important to improve the quality of the statistical analysis, but it is also essential to discuss the results, when we start to examine how the local properties of the sample compare with the local sequence. It is a step that should be included in all SAXS studies on polymer chains. In theory, if the chains do not have any intrinsic asymmetry, they could be tagged by an additional short domain at one end, chosen for its properties to be easily identified in the SAXS data, and so allow their orientation.

The analysis can then be extended to detect other properties of the sample. In Section 3, we show how the dihedral angles vary along the sequence, and we also calculate the probability $P(n, |\theta|)$ showing, for each site, the distribution of the bending angles among the 1000 best conformations, and similarly for $P(n, |\varphi|)$. These figures are very helpful to obtain a view of the likely conformations of the samples in solution and to try to understand why some sequences may be favourable for protein binding. Because long polymers, such as DNA, fluctuate extensively in solution, a statistical analysis of their most likely conformations is crucial to understanding biological phenomena. It provides information that usefully complements high-resolution structures of protein–DNA complexes, which often only show a snapshot of the final conformational state after binding.

3. Results and discussion

3.1. Small-angle X-ray scattering data

The chromatograms computed with the program *CHROMIXS* (Panjkovich & Svergun, 2018) indicated the presence of one single main peak eluting in a similar position for all of the duplexes, with some small shoulders on either side of the peak (Supplementary Fig. S1). All eluted peaks had 260:280 nm UV absorbance ratios consistent with that of pure dsDNA (Supplementary Fig. S1). Each dsDNA peak had some instability in predicted R_g values across the eluted peak, possibly because of neighbouring peaks. However, frames were selected for analysis from regions where the R_g was stable across consecutive frames within approximately 5 Å or less (Supplementary Fig. S1). In the case of GAGE6, the R_g across the chosen frames varied from 45.3 to 50.78 Å, with an average of 48.08 Å. For GAGE6_1 the R_g varied from 49.9 to 52.9 Å, with an average of 52.0 Å. For GAGE6_2 the R_g of the chosen frames ranged from 53.9 to 54.0 Å, with the R_g in the surrounding frames varying from 51.5 to 54.0 Å. In the case of GAGE6_3, the variation in the predicted R_g across the chosen frames was minimal, ranging from 49.3 to 50.6 Å, with an average of 50.1 Å.

The calculation of the molecular weight of the frames across the chromatography peak worked well for GAGE6_1, GAGE6_2 and GAGE6_3; however, it was unsuccessful for GAGE6, perhaps due to the fact that this experiment is slightly noisier than the others. As shown in Supplementary Fig. S2, the molecular weights of the frames across the peaks are in agreement with the expected molecular weight of a dsDNA duplex of this size (∼37 kDa). This demonstrates that at least in the case of three of the four chromatograms, the samples are clearly monodisperse. Despite this calculation not working for the GAGE6 sample, the R_g values of the chosen frames for the GAGE6 experiment are in close agreement, demonstrating that this sample is also likely to be monodisperse.

All scattering shown as $\log_{10}(I)$ versus $\log_{10}(q)$ plots demonstrated good-quality data with a plateau in the Guinier region (Fig. 5). When constrained to $qR_g \lesssim 1$ (suitable for elongated molecules), the Guinier fits had high fidelity scores and acceptable distributions of residuals. The fits also provided R_g values similar to the expected value for a straight 60 bp dsDNA duplex as simulated with *CRYSOL* (R_g = 56.10 Å; Fig. 5). The distance distribution functions of the various oligonucleotides from this study highlight some robust features across each variant. Depending on the duplex the shape of the $P_{exp}(r)$ function changes, with characteristic bumps in the GAGE6, GAGE6_1 and GAGE6_2 functions which then almost disappear in GAGE6_3, as shown in Figs. 6, 7, 8, 9 and 10. The D_{max} values for all functions were within a ±10% range of the theoretical D_{max} of a 60 bp duplex. Supplementary Fig. S5 shows that deviation of D_{max} from the optimal value given by our procedure only leads to small variations in $P_{exp}(r)$. Additionally, Supplementary Fig. S6 shows that despite variation of the *GNOM* regularization parameter the bumps in our respective $P_{exp}(r)$ functions persist. The R_g values derived from the individual $P_{exp}(r)$ functions were more or less consistent with those from the Guinier approximation. We noted minor inconsistencies between $P_{exp}(r)$ and Guinier-derived R_g for GAGE6 and GAGE6_1. The details concerning SAXS parameters are summarized in Table 1.

3.2. Sequence GAGE6

Let us start with the GAGE6 sequence, which is the leading sequence of the series since it is a native promoter sequence that is bound by SFPQ. Fig. 6(*a*) compares the experimental pair-distance distribution function $P_{exp}(r)$ with $P(r)$ calculated from the 1000 best conformations saved after running Monte Carlo simulations to generate conformations with the polymer model. The agreement between the two distributions is evaluated by computing

$$\chi^2 = \frac{1}{N_r - 1} \sum_{1}^{N_r} \left[\frac{P(r_j) - P_{exp}(r_j)}{\sigma_j} \right]^2, \qquad (6)$$

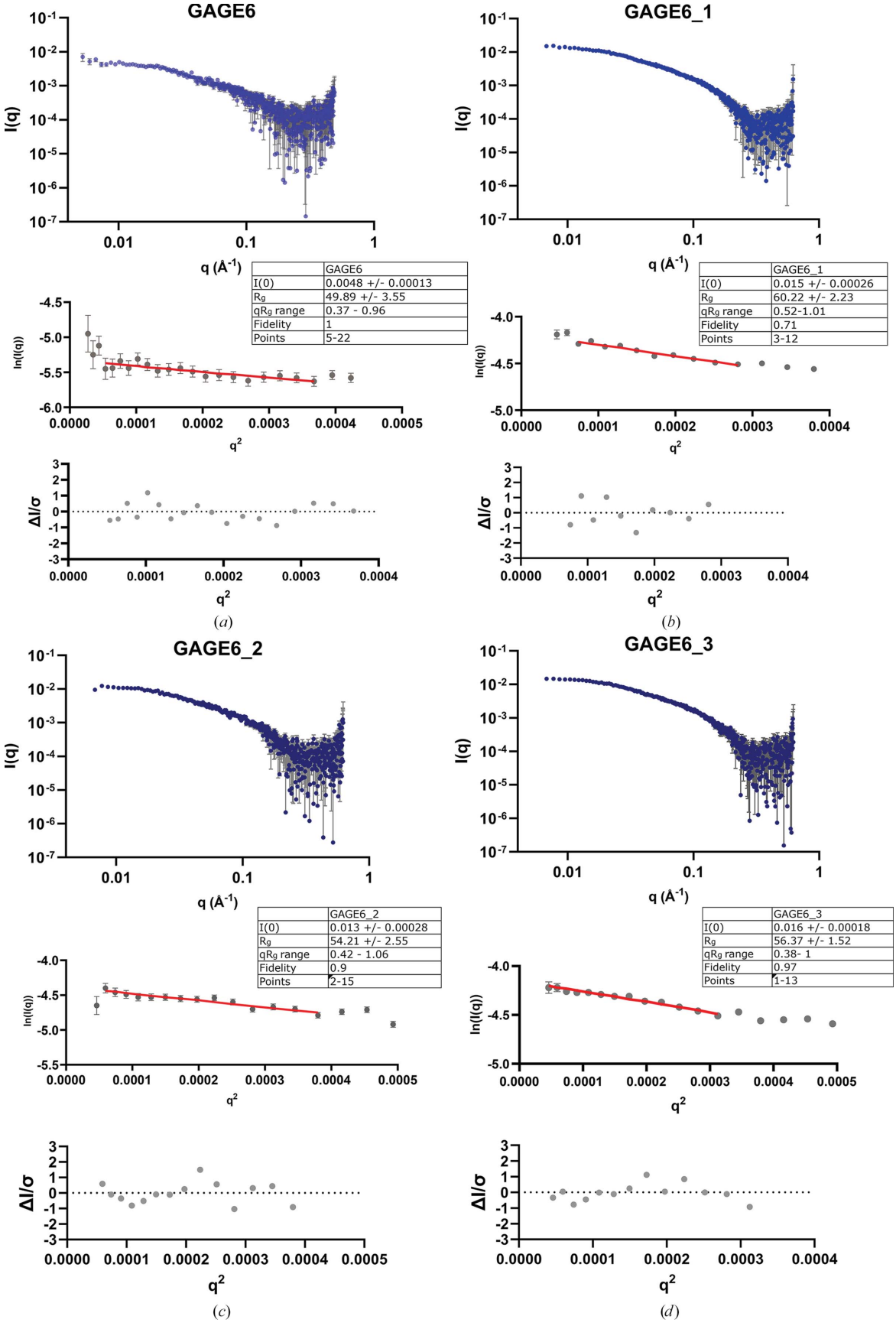


Figure 5
Scattering data and Guinier fits of the various GAGE6 oligonucleotides: $I(q)$ versus q plot, Guinier plot (ln[$I(q)$] versus q^2) and Guinier residual plots for (*a*) GAGE6, (*b*) GAGE6_1, (*c*) GAGE6_2 and (*d*) GAGE6_3. Guinier fit parameters and results are tabulated in the figure.

where r_j are the points where $P_{exp}(r)$ has been calculated, σ_j is the experimental error at these points (estimated by the *GNOM* program) and N_r is the number of calculation points. The summation is restricted to $r_j > 25$ Å because the polymer model cannot describe the internal structure of DNA, which dominates $P_{exp}(r)$ for $r < 25$ Å as discussed above (Section 2.5.2). The curve deduced from the selected conformations of the polymer model (Fig. 6*a*) provides a very good description of the experimental curve in the whole domain where the polymer model is valid, indicating in this instance that the polymer model ensemble can very accurately reproduce $P_{exp}(r)$.

Fig. 6(*b*) shows the statistics of the bending angles $|\theta_n|$ and dihedral angles $|\varphi_n|$ with their averages over conformations and their standard deviations. For the bending angles this figure is analogous to Fig. 4 except that the conformations have been oriented as described in Section 2. Only one of the double maxima that were forming a mirror image with respect to the middle of the sequence has subsisted, which confirms that the symmetric pattern was created through symmetry inherent in the SAXS measurement. The value $\Delta s/s$ marked on the figure corresponds to $(s' - s)/s$, where s and s' were the local sums of the bending angles in the two mirror domains used to orient the conformations. Their differences were

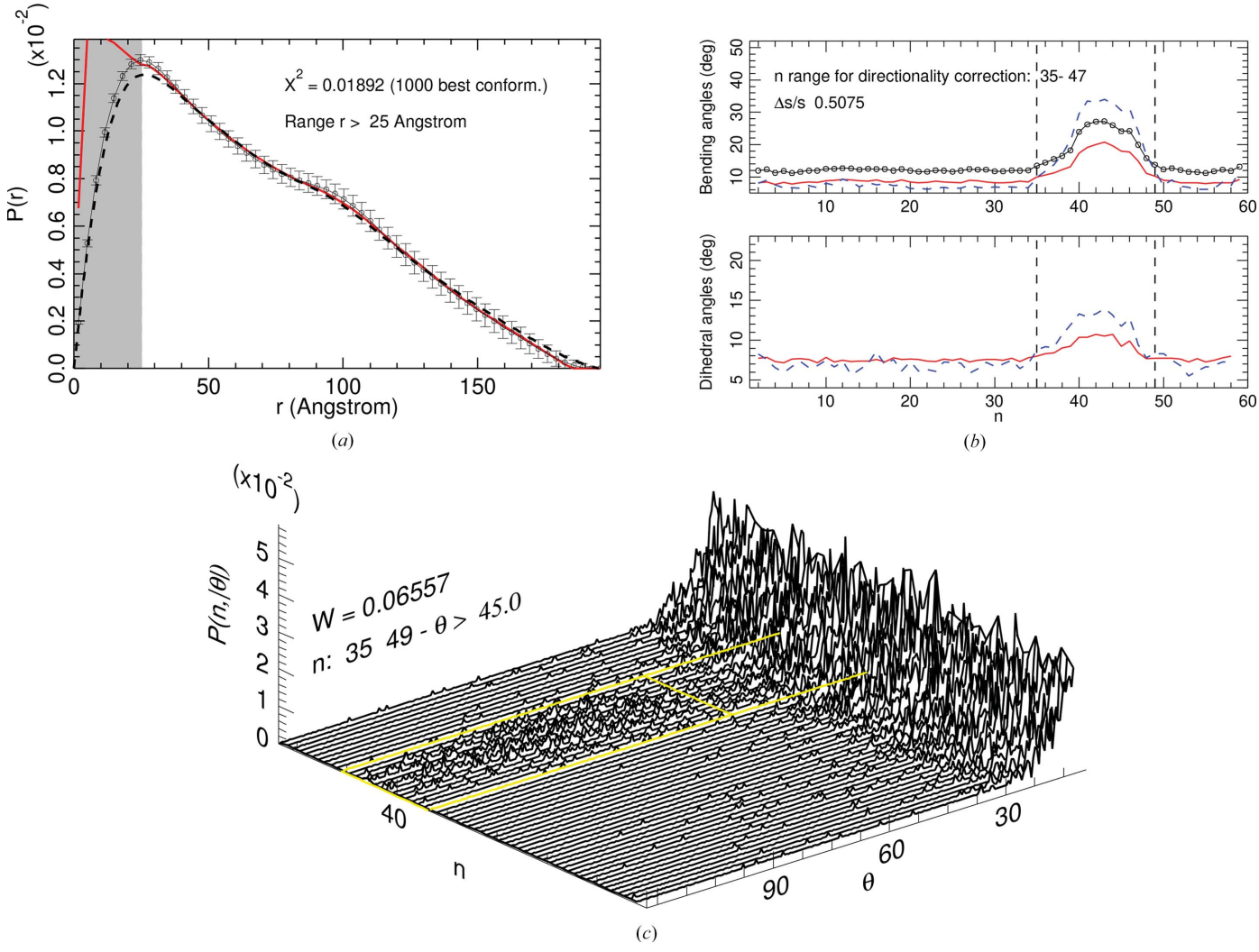


Figure 6
Sequence GAGE6. (*a*) Comparison between $P_{exp}(r)$ calculated with *AutoGNOM*-5 (small circle points and thin black curve with error bars) and $P(r)$ calculated for the 1000 best conformations generated theoretically (thick red curve). The grey region, for $r < 25$ Å, is the low-r domain in which the model cannot describe the internal structure of the double helix. The thick black dashed line shows the function $P_{exp}(r)$ obtained with *AutoGNOM*-4 (see Section 2.4). (*b*) Upper part: statistics of the absolute value of the local bending angles $|\theta_n(i)|$ over the 1000 best saved conformations for the GAGE6 sample after orientation of the conformations as discussed in Section 2. The red full line shows the average over i of $|\theta_n(i)|$ versus n. The black line with circles shows the same average, limited to the bending angles which are above $|\theta| = 5°$, and the dashed blue curve shows the standard deviation of $|\theta_n(i)|$. Lower part: statistics of the absolute values of the dihedral angles. The red full line shows the average over i of $|\varphi_n(i)|$ versus n and the the dashed blue curve shows their standard deviation. (*c*) Three-dimensional plot of the normalized probability distribution function $P(n, |\theta|)$. The data are not shown for small bending angles, for which the probabilities are the highest, to better show the data at larger θ. The value W marked on the figure shows the fraction of bending angles $|\theta_n|$ above the value shown below W within the domain indicated by the two values of n.

significant, suggesting that one of the two mirror images was fictitious, which is confirmed by the figure as one of them has been fully erased in the orientation step.

The three-dimensional plot in Fig. 6(*c*) shows the probability $P(n, |\theta|)$ of the bending angles $|\theta_n|$ for each site n along the sequence. These distributions are normalized

$$\int_0^{\pi} P(n, |\theta|)\, \mathrm{d}|\theta| = 1 \quad \text{for each } n, \tag{7}$$

which allows us to calculate the fraction W of the bending angles $|\theta_n|$ above some value in a given domain. The result $W = 0.06557$ is shown in Fig. 6(*c*) for $35 \leq n \leq 49$ and a lower value of θ of 45°. Thus, although most of the bending angles are well below 30°, in some regions of the sequence about 6% of the bending angles are above 45°. The statistics clearly show that one region of the sequence is special, with a much higher flexibility and larger bending than elsewhere. The colour view of the three-dimensional probability $P(n, |\theta|)$ (Fig. 6*c*; see Supplementary Fig. S9) suggests that the large bending angles for $35 \leq n \leq 49$ are separated from the bulk of smaller bending angles, as if they result from some permanent bend in this region and not simply from large-amplitude thermal fluctuations. As this part of the probability distribution only concerns 6% of the bending angles, the resolution of the analysis of the SAXS data does not however allow us to make a definitive statement on this point.

We would like to stress that these statistical results which provide precise information on the local properties of the sample are extracted from a probability distribution function $P_{\rm exp}(r)$ which looks rather featureless, although it results from a large ensemble of pair distances. It is remarkable that this smooth-looking, almost straight function actually encodes so much information. Now that we have oriented the conformations we can also compare the statistical results with the actual sequence of the sample. There is still an unknown point: should the sequence listed in Fig. 1 be oriented from 1 to 60 or from 60 to 1?

The GAGE6 sequence is asymmetric, with large AT-rich domains at the 3′ end, while the 5′ end is more GC-rich. The AT base pairs, which are linked by only two hydrogen bonds, are generally more flexible than the GC pairs linked by three hydrogen bonds because their weaker bonding allows greater freedom in their local conformations. This effect is also enhanced by easier fluctuational openings, even at room temperature (Theodorakopoulos & Peyrard, 2012). Therefore, it seems reasonable to assume that the $n = 60$ side of our diagrams corresponds to the 3′ end. With this assignment, we can notice that the domain $n = 35 \ldots 47$ that we used for the orientation step is almost completely made of AT pairs with only two, separate, GC pairs inside it. Therefore, it is not surprising to find that this domain is highly flexible, as found in the analysis of the SAXS data. Thus, the results of our analysis of the SAXS data suggest that this polymer-modelling pipeline is able to detect specific features at the scale of a few base pairs. Its validity is strengthened by the results on the fluctuations of the dihedral angles, which show a maximum in their standard deviations which coincides with the maximum of the bending angles, although it is slightly broader. In the linear chain of joined segments that we use to analyse the data there is no geometrical constraint which links bending and twist. However, in DNA there is such a constraint due to the double-helical structure. It is easy to check with a mechanical device such as two wires twisted together to make a helix that a large amount of bending tends to locally untwist the device, simply by geometrical effects. The fit of our polymer model to $P_{\rm exp}(r)$ detects this phenomenon although it is not built into our hypothesis.

However, although the sequence strongly contributes to determining the local flexibility of the sample, the results also point to subtle effects associated with some cooperativity along the double helix. In the GAGE6 sequence, sites 38–44 make a large continuous domain with only AT pairs and therefore it is not surprising to find higher bending angles in this region, as pointed out above. However, the GAGE6 sequence also has a four-AT-pair domain at sites 28–31 which are not within the region where we detect larger bending angles. Similarly, the five-AT-pair domain at sites 50–54 also shows fairly low bending angles.

The sequence provides a hint on the local flexibility of DNA, but analysis of the SAXS data suggests that experiments which are sensitive to the conformation of DNA tell us more. This may have many biological implications. Within cells and viruses, DNA has to be tightly bent or kinked at many sites to allow compact packaging of large amounts of genetic material (Widom, 1984). Many studies have tried to address the question of whether DNA bending arises from metastable bent structures or more flexible 'joints' in dsDNA. Previously solved structures of dsDNA AT-rich dodecamers (Stefl *et al.*, 2004; Hizver *et al.*, 2001) suggest that bent duplexes are stable and are maintained in a bent conformation by several molecular features. This notion was supported by a fluorescence polarization experiment (Chirico *et al.*, 2001) that also argued for a permanent static bend in some AT-rich dsDNA sequences. A review by Harteis & Schneider (2014) of the structure of DNA summarized that whilst AT tracts can introduce an intrinsic bend, they in fact somewhat increase the rigidity of such sequences due to certain stabilizing effects that accommodate the bending of DNA: propellor twisting of bases, a narrow minor groove and a wider major groove (Stefl *et al.*, 2004; Hizver *et al.*, 2001), as well as non-Watson–Crick hydrogen bonds forming down the major groove. Rather interestingly, the bending of AT-rich DNA was also shown to be sensitive to the order of the bases, with A4T4 versus T4A4 steps having different effects: the former causing a global bend in the helix and the latter forming a straighter helix (Stefl *et al.*, 2004). DNA has been considered to act more flexibly when a 'kink', defined as the local unstacking of base pairs causing a change in the orientation of the helix, appears (Harteis & Schneider, 2014). Pyrimidine–purine (TA and CA) steps reportedly exhibit the lowest base-stacking energy and can behave more like flexible hinges that can cause DNA bending.

In addition to these structural effects, further studies have also examined the fluctuational opening of DNA base pairs or

the formation of 'bubbles' and whether these could be a source of flexibility in DNA. Guéron *et al.* (1987) described the probability of the fluctuational opening of base pairs as low via imino-proton exchange. More recently, Theodorakopoulos & Peyrard (2012) reviewed this concept and theorized that below the melting point of a duplex, an increase in temperature towards physiological temperature may increase the opening of base pairs enough for them to act as flexible hinges in dsDNA, which may have wider implications for DNA structure and recognition by proteins.

Alongside the fluctuational opening of the duplex, further unwinding of DNA regions is also important. Taking a simple physical model of two interwound strands and bending it sufficiently causes unwinding of the two strands at the point of bending. Recently, Chandrasekhar *et al.* (2024) demonstrated via a minicircle assay that DNA can locally unwind in response to bending. This has implications for gene-regulatory and DNA-packaging elements, as the bending of DNA by any of the aforementioned mechanisms may also contribute to the unwinding of the duplex.

In a SAXS experiment, it can be complicated to delineate between interrelated effects such as stable, rigid/bent poly-AT structures, flexibly kinked DNA as a result of unstacked base pairs, the unwinding of strands or the fluctuational opening of base pairs. This is particularly true given that SAXS is, relatively speaking, a low-resolution structural technique. However, having shown that we can quantitatively detect the bending of DNA in solution, we hypothesized that such events could be described by the complementary study of a series of related sequences derived from GAGE6.

3.3. Sequences GAGE6_1, GAGE6_2 and GAGE6_3

This series of three sequences was derived from GAGE6 by incrementally removing its longest AT-rich domains and replacing them with GC-rich segments. This allowed us to test two things:

(i) to check how the analysis of the SAXS data detects these local changes and how they influence the conformations of the samples, and

(ii) to try and understand the properties of the GAGE6 sequence that allow it to bind SFPQ. The first sequence of the series, GAGE6_1, only differs from GAGE6 by four bases at sites 28–31. The TTTT sequence in GAGE6 has been replaced by GCGC in GAGE6_1.

Fig. 7 shows the equivalent data for the GAGE6_1 variant to that in Fig. 6 for GAGE6. Fig. 7(*a*) shows that although the change in the sequence is very localized, the shape of $P_{\rm exp}(r)$ is significantly modified: the single bump around 100 Å in GAGE6 splits into two lumps around 70 and 125 Å for GAGE6_1. Our analysis shows that this difference is due to a large change in the distribution of θ angles (Fig. 7*b*). The broad domain with large θ angles covering sites 36–46 in GAGE6 is now narrower and restricted to sites 44–52 only, but is also much sharper, as the maximum of $\theta_n(i)$ reaches 36° instead of about 20° for GAGE6. The dihedral angles show the same behaviour, with large twist fluctuations being restricted to the same narrow domain.

Fig. 7(*c*) confirms the presence of large bending angles in a narrow domain and the integration of $P(n, |\theta|)$ for $41 \leq n \leq 52$ and $\theta > 45°$ shows that the fraction W of bending angles $|\theta_n|$ above 45° has drastically increased from 6.4% to 12.6%. The GAGE6_1 sequence therefore has a high probability of being sharply bent in a narrow region. It is interesting to notice the qualitative analogy between the experimental pair-distance distribution function for GAGE6_1 and the double-bump simulated $P(r)$ (Fig. 2) obtained with an atomistic model of sharply bent DNA. This strengthens the notion that $P(r)$ has a high sensitivity to molecular conformations, particularly bending, with sharp bending generating a function with two bumps.

In GAGE6_2, in addition to replacing the AT base pairs between positions $28 \leq n \leq 31$ with GC pairs, the AT pairs at positions $38 \leq n \leq 44$ have now also been replaced with GC pairs. As expected, the change in properties of the sample with respect to GAGE6 is even more drastic. Fig. 8(*a*) shows that $P_{\rm exp}(r)$ has sharper lumps, slightly moved towards larger values of r. The error bars given by *GNOM* in the calculation of $P_{\rm exp}(r)$ from $I(q)$ are larger and the fit of this pair-distance distribution function by the simple model is less accurate, although it remains within the error bars almost everywhere.

Fig. 8(*b*) shows that the distribution of large bending angles along the sequence, as well as the domains where the dihedral angles have large standard deviations, have qualitatively changed. Fig. 8(*b*) appears to be very consistent with the sequence of GAGE6_2. Its largest AT domains are $10 \leq n \leq 13$ and a large domain from $n = 50$ to $n = 57$ which is only interrupted by a single GC pair at position $n = 55$. This agrees well with the small increase of bending and twist fluctuations seen around $n = 10$ in Fig. 8(*b*) and the larger bending angles and twist fluctuations in the range $44 \leq n \leq 57$ split into two peaks by a dip at sites 52 and 53 which could correspond to the GC pair at site 55 within the resolution given by the analysis of the SAXS data. The bending angles $|\theta_n(i)|$ in this domain are however much smaller than those in the sharp bending domain of GAGE6_1. This qualitative change from GAGE6_1 may nevertheless seem surprising if we think that the only difference in sequence between GAGE6_1 and GAGE6_2 is the elimination of an AT-rich domain between positions $38 \leq n \leq 44$. Both GAGE6_1 and GAGE6_2 have a large AT-rich domain ($50 \leq n \leq 57$), but in GAGE6_2 this region causes a smaller bending effect than in GAGE6_1. This points out again that although sequence and flexibility are correlated, the preferred conformations of a DNA sequence in solution are not only determined by the local sequence; collective effects at larger distances are also important. The probability distribution of the bending angles $P(n, |\theta|)$ shown in Fig. 8(*c*) and Supplementary Fig. S9 are somewhat similar to the data from Fig. 7: the fraction of large bending angles $|\theta_n|$ in the range $44 \leq n \leq 57$ is of the order of 9.6%, suggesting that GAGE6_2 is also sharply bent in a narrow region.

However, the results shown in Fig. 8(*a*) raise questions because the fit of the polymer model to the $P_{\rm exp}(r)$ calculated

with *AutoGNOM*-5 looks rather poor. We know that the model is oversimplified to fully describe DNA, but for $r > 25$ Å it should nevertheless provide a better description of the experiment because we have extensively explored the conformational space of the sample by generating more than 4×10^8 conformations and selecting the best 1000 for the analysis. This failure suggests that the difficulty could come from $P_{exp}(r)$ itself. At this point we would like to stress the importance of choosing a model which has realistic bending and torsional properties to make sure that the conformations which enter into our fits are actually accessible to DNA molecules. A purely random model, such as the Gaussian model which can adopt any conformation at no energy cost, might be able to provide a good fit to $P_{exp}(r)$, but with unphysical DNA conformations. In Section 2.4 we discussed that for sequence GAGE6_2 the determination of the optimal regularization parameter α by *GNOM* is particularly hard. *GNOM*-5 converges to $\alpha = 0.01084$, while *GNOM*-4 converges to $\alpha = 1.31$ and generates a much smoother $P_{exp}(r)$ with smaller estimated error bars. We conjectured that for a given sample, the functions $P_{exp}(r)$ for different values of α could contain almost the same information on the physical properties of the DNA sample. To check this conjecture, we repeated with the *AutoGNOM*-4 $P_{exp}(r)$ functions the same analysis initially performed with those derived using *AutoGNOM*-5. The results are shown in Fig. 9.

The top panel of Fig. 9 shows that the model gives an almost perfect fit of *AutoGNOM*-4 $P_{exp}(r)$ which is much smoother

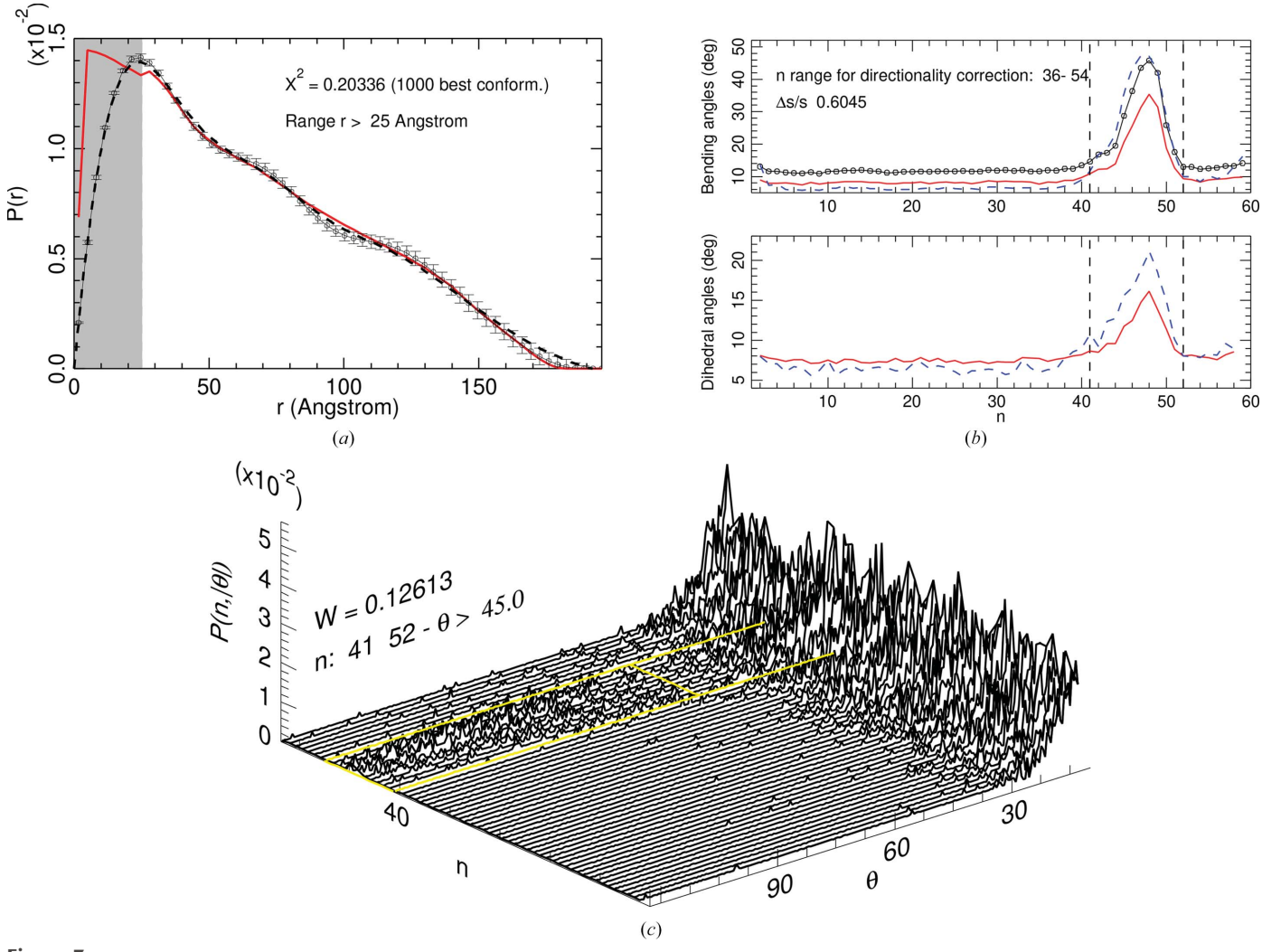


Figure 7
Sequence GAGE6_1. (*a*) Comparison between $P_{exp}(r)$ calculated with *AutoGNOM*-5 (small circle points and thin black curve with error bars) and $P(r)$ calculated for the 1000 best conformations generated theoretically (thick red curve). The grey region, for $r < 25$ Å, is the low-r domain in which the model cannot describe the internal structure of the double helix. The thick black dashed line shows the function $P_{exp}(r)$ obtained with *AutoGNOM*-4 (see Section 2.4). (*b*) Upper part: statistics of the absolute value of the local bending angles $|\theta_n(i)|$ over the 1000 best saved conformations for the GAGE6_1 sample, after orientation of the conformations as discussed in Section 2. The red full line shows the average over i of $|\theta_n(i)|$ versus n. The black line with circles shows the same average, limited to the bending angles which are above $|\theta| = 5°$, and the dashed blue curve shows the standard deviation of $|\theta_n(i)|$. Lower part: statistics of the absolute values of the dihedral angles. The red full line shows the average over i of $|\varphi_n(i)|$ versus n and the the dashed blue curve shows their standard deviation. (*c*) Three-dimensional plot of the normalized probability distribution function $P(n, |\theta|)$. The data are not shown for small bending angles, for which the probabilities are the highest, to better show the data at larger θ. The value W marked on the figure shows the fraction of bending angles $|\theta_n|$ above the value shown below W within the domain indicated by the two values of n.

than *AutoGNOM*-5 $P_{exp}(r)$. In comparing Figs. 8 and 9, one should keep in mind that the definition of χ^2 (equation 6) compares the discrepancy between $P_{exp}(r)$ and the theoretical $P(r)$ with the estimated error on $P_{exp}(r)$. As the error bars on *AutoGNOM*-4 $P_{exp}(r)$ are much smaller than those on *AutoGNOM*-5 $P_{exp}(r)$, similar values for χ^2 in both figures imply that the discrepancies are much smaller in Fig. 9. The bottom panel of Fig. 9 shows that although the two functions $P_{exp}(r)$ look radically different, they provide almost the same results concerning the propensity of the DNA sample to bend in the 44–57 domain, which is the AT-rich domain. However, the *AutoGNOM*-4 $P_{exp}(r)$ misses the large degree of bending occurring near n = 56–57 which is detected by the *AutoGNOM*-5 $P_{exp}(r)$. Although it cannot be formally excluded that the bending detected by *AutoGNOM*-5 $P_{exp}(r)$ in the vicinity of an AT pair is an artefact, it is likely that the high value of α derived by *AutoGNOM*-4 leads to the over-regularization of this function, which smoothes out this local effect. Nevertheless, it is remarkable that two functions $P_{exp}(r)$ which look so different at a first glance encode the same information about the bendability of a sample near one of its ends. This supports our conjecture and implies that a perfect refinement of $P_{exp}(r)$, which may be difficult, is not an absolute requirement to analyse SAXS data with a polymer model as proposed in this work.

In GAGE6_3, the last of the series of sequences derived from GAGE6, the last large AT-rich domain that spanned positions $50 \leq n \leq 54$ has been replaced entirely by GC base

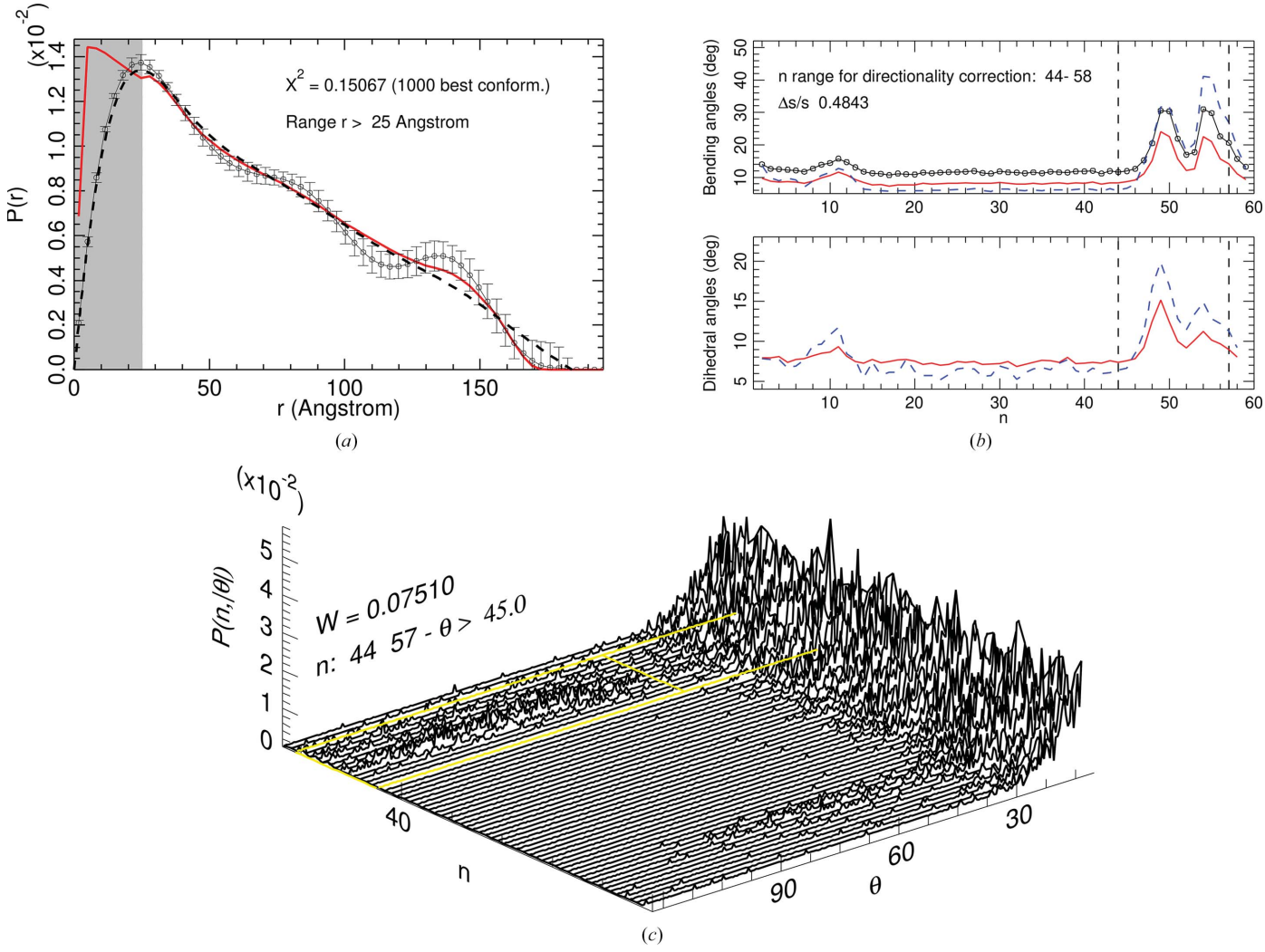


Figure 8
Sequence GAGE6_2. (*a*) Comparison between $P_{exp}(r)$ calculated with *AutoGNOM*-5 (small circle points and thin black curve with error bars) and $P(r)$ calculated for the 1000 best conformations generated theoretically (thick red curve). The grey region, for $r < 25$ Å, is the low-r domain in which the model cannot describe the internal structure of the double helix. The thick black dashed line shows the function $P_{exp}(r)$ obtained with *AutoGNOM*-4 (see Section 2.4). (*b*) Upper part: statistics of the absolute value of the local bending angles $|\theta_n(i)|$ over the 1000 best saved conformations for the GAGE6_2 sample, after orientation of the conformations as discussed in Section 2. The red full line shows the average over i of $|\theta_n(i)|$ versus n. The black line with circles shows the same average, limited to the bending angles which are above $|\theta| = 5°$, and the dashed blue curve shows the standard deviation of $|\theta_n(i)|$. Lower part: statistics of the absolute values of the dihedral angles. The red full line shows the average over i of $|\varphi_n(i)|$ versus n and the the dashed blue curve shows their standard deviation. (*c*) Three-dimensional plot of the normalized probability distribution function $P(n, |\theta|)$. The data are not shown for small bending angles, for which the probabilities are the highest, to better show the data at larger θ. The value W marked on the figure shows the fraction of bending angles $|\theta_n|$ above the value shown below W within the domain indicated by the two values of n.

pairs. Fig. 10(*a*) shows that $P_{exp}(r)$ has lost its two-lump structure and conforms more, in comparison to the other experimental functions, to the simulated shape of a straight duplex (Fig. 2). The disagreement between the polymer model fit and $P_{exp}(r)$ at large values of r here may be a result of the former not accounting for hydration-shell effects whilst the latter does. Although the large AT-rich regions of the $n > 30$ range that we used to orient the conformations for previous sequences are no longer present, the sequence still has asymmetry. The domain $10 \leq n \leq 26$ has several short AT sequences, while its mirror image $35 \leq n \leq 49$ now only has three AT pairs and is likely to be less flexible, being systematically enriched in GC content. This is confirmed by the sums s and s' of $|\theta_n|$ in these two regions, which differ systematically from each other and can be used to orient the conformations with respect to the sequence. Fig. 10(*b*), which does not display the symmetry expected for unoriented SAXS data, shows that the orientation procedure described in Section 2 is again valid for GAGE6_3. Fig. 10(*b*) shows a broad domain with large bending angles with a sharp maximum at position $n = 26$ which coincides with two AT pairs in the middle of a GC sequence. The three-dimensional plot of the probability $P(n, |\theta|)$ plotted in Fig. 10(*c*) and Supplementary Fig. S9 shows larger bending angles in the domain $17 \leq n \leq 35$, but the integrated weight of the probability $P(n, |\theta|)$ for angles exceeding 45° in this region shows that only 3.8% of the angles correspond to large bending. As suggested by its sequence, which comprises mostly GC base pairs and only a few small domains with AT pairs, not wider than three consecutive sites, the GAGE6_3 sequence has mostly straight conformations.

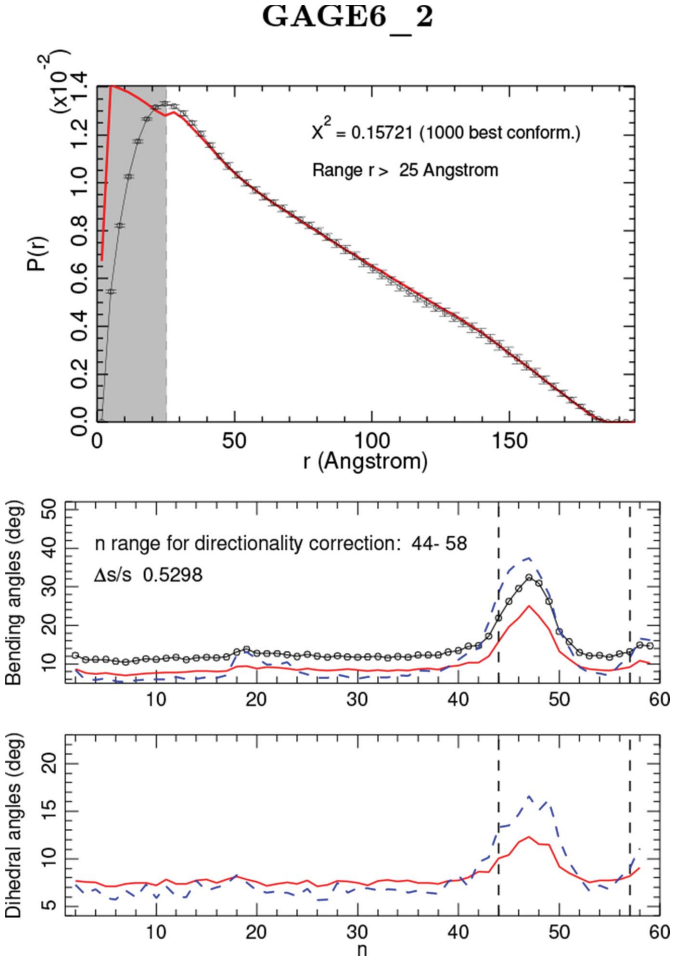


Figure 9
Sequence GAGE6_2. Top panel: comparison of *AutoGNOM*-4 $P_{exp}(r)$ (small circle points and thin black curve with error bars) and $P(r)$ given by the 1000 best conformations of the polymer model (thick red curve). The grey region, for $r < 25$ Å, is the low-r domain in which the model cannot describe the internal structure of the double helix. Bottom panel, upper part: statistics of the absolute value of the local bending angles $|\theta_n(i)|$ over the 1000 best saved conformations for the GAGE6_2 sample, after orientation of the conformations as discussed in Section 2. The red full line shows the average over i of $|\theta_n(i)|$ versus n. The black line with circles shows the same average, limited to the bending angles which are above $|\theta| = 5°$, and the dashed blue curve shows the standard deviation of $|\theta_n(i)|$. Bottom panel, lower part: statistics of the absolute values of the dihedral angles. The red full line shows the average over i of $|\varphi_n(i)|$ versus n and the the dashed blue curve shows their standard deviation. The vertical dashed lines show the boundaries of the domain where large bending was observed in the analysis of *AutoGNOM*-5 $P_{exp}(r)$.

Supplementary Fig. S10 shows the results of the analysis of the functions $P_{exp}(r)$ obtained with *AutoGNOM*-4 for the four samples (see Supplementary Fig. S6). It can be compared with Figs. 6, 7, 9 and 10 showing the results obtained with *AutoGNOM*-5. It confirms our conclusions about the domains prone to extra flexibility and bending in each sequence. However, it appears that large regularization parameters α, giving smoother $P_{exp}(r)$, may lead to lower amplitudes of the calculated bending and flexibility in the DNA sites belonging to those domains and, in some cases, loss of fine structures. Our results suggest that for an analysis with a model which properly describes DNA bending and torsional rigidity, and therefore avoids unphysical conformations, as the choice of the optimal regularization parameter is difficult, favouring small regularization parameters α may be a good choice because it leads to $P_{exp}(r)$ functions which encode more structural details.

The study of a series of four DNA sequences, including GAGE6 and three others obtained by gradually replacing AT-rich domains with GC tracts, is interesting because the unique structural features of bent AT-tract-containing duplexes can be important for the recognition of dsDNA by proteins (Harteis & Schneider, 2014). Typically, the major groove in dsDNA has the highest potential for protein–base recognition as the functional groups of the four bases are accessible (Harteis & Schneider, 2014). In the case of bent AT-tract-containing structures, the major groove shows an even higher potential for base recognition, as the groove is often widened by bending (Harteis & Schneider, 2014; Stefl *et al.*, 2004). Bending likely serves to enhance sequence-specific recognition of DNA by proteins in many contexts. A simultaneous effect in AT-tract-containing dsDNA is the narrowing of the minor groove (Stefl *et al.*, 2004). The narrower groove places elements of the backbone much closer to one another, resulting in an enhanced electrostatic effect that can facilitate structure-specific rather than sequence-specific recognition by proteins (Ferrari *et al.*, 1992; Rohs *et al.*, 2009).

It is possible that SFPQ recognizes the GAGE6 oligonucleotide through an interaction of the RGG domains with

either a wider major groove or a more pronounced electrostatic interaction with the backbone due to the narrower minor groove. The latter is an interesting possibility given that RGG motifs are the putative DNA-binding regions of SFPQ, and arginine is the most common residue found to interact with narrow minor grooves (Rohs *et al.*, 2009). Another possibility is that the fluctuational opening of the 'TTTATTT' region in the GAGE6 oligonucleotide, which was the approximate region where the majority of bending was detected in our analysis, plays a role in protein binding. This effect may be more pronounced at a higher temperature (Theodorakopoulos & Peyrard, 2012) or inside biological condensates, which through multivalent competitive interactions can reportedly cause the melting of dsDNA (Nott *et al.*, 2016). The truncation of the GAGE6 sequence from 60 to 40 bp reduced the binding of SFPQ from 84% to 45% as calculated via a gel-shift assay (Wang *et al.*, 2022). This truncation resulted in the 'TTTATTT' domain being cut in half. This suggests that the last 20 bp of the sequence may provide the region required for the enhanced binding of SFPQ. However, as functional aggregation of the protein is a key component of binding (Lee *et al.*, 2015; Koning *et al.*, 2025) it is possible that 40 bp was simply too short a length to allow multiple units of SFPQ to have several points of contact with the DNA.

The idea that certain sites in dsDNA offer less resistance to structural deformation and so a lower energetic penalty to protein–DNA complexes that deform rigid DNA is also

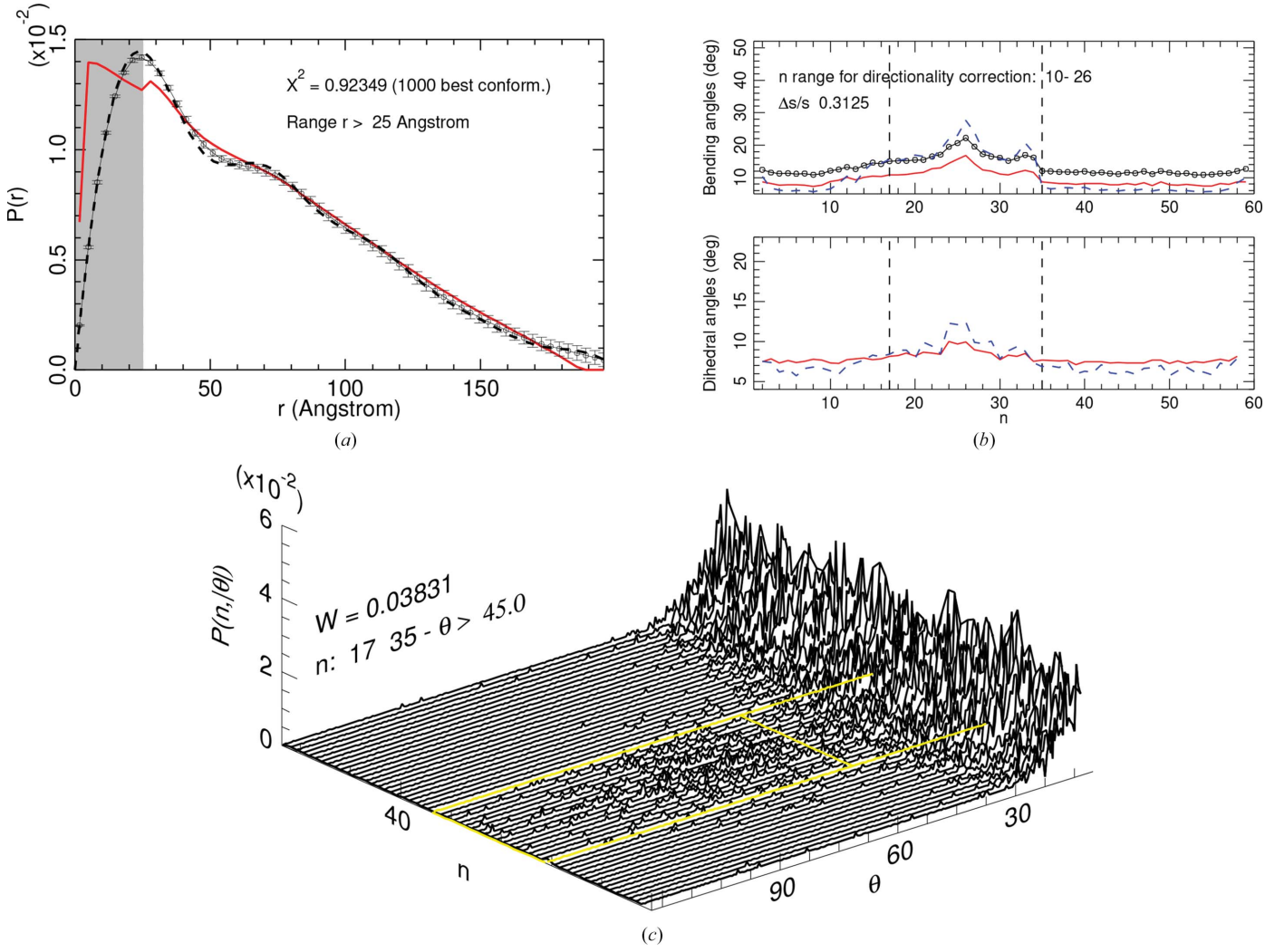


Figure 10
Sequence GAGE6_3. (*a*) Comparison between $P_{exp}(r)$ calculated with *AutoGNOM*-5 (small circle points and thin black curve with error bars) and $P(r)$ calculated for the 1000 best conformations generated theoretically (thick red curve). The grey region, for $r < 25$ Å, is the low-r domain in which the model cannot describe the internal structure of the double helix. The thick black dashed line shows the function $P_{exp}(r)$ obtained with *AutoGNOM*-4 (see Section 2.4). (*b*) Upper part: statistics of the absolute value of the local bending angles $|\theta_n(i)|$ over the 1000 best saved conformations for the GAGE6_3 sample, after orientation of the conformations as discussed in Section 2. The red full line shows the average over i of $|\theta_n(i)|$ versus n. The black line with circles shows the same average, limited to the bending angles which are above $|\theta| = 5°$, and the dashed blue curve shows the standard deviation of $|\theta_n(i)|$. Lower part: statistics of the absolute values of the dihedral angles. The red full line shows the average over i of $|\varphi_n(i)|$ versus n and the the dashed blue curve shows their standard deviation. (*c*) Three-dimensional plot of the normalized probability distribution function $P(n, |\theta|)$. The data are not shown for small bending angles, for which the probabilities are the highest, to better show the data at larger θ. The value W marked on the figure shows the fraction of bending angles $|\theta_n|$ above the value shown below W within the domain indicated by the two values of n.

important for many protein–DNA interactions. Naturally, if a protein is required to bend DNA as part of its function, a pre-bent or flexible substrate would dramatically reduce the amount of energy required to bend DNA. This concept has been reviewed by Harteis & Schneider (2014) and some examples are given below.

ECOR1 significantly bends DNA at its recognition site 'GAATTC' as part of cleavage reactions (Widom, 1984). The degree of natural bending at this target site likely lowers the energy required for binding, as bending of the unbound DNA can place it close to a position that already resembles the bound state. Gartenburg & Crothers (1988) examined the degree of DNA bending caused by the catabolite activator protein (CAP) in response to different mutated DNA targets. DNA bending in regions flanking the binding site enhanced the affinity of the CAP–DNA complex by creating a larger and more complex interface. HMG-box domains in transcriptional regulatory proteins bind DNA. In general, all of the binding sites for these proteins are AT-rich, and the protein SRY has been shown to induce a sharp bend when binding to the sequence 'AACAAG' (Ferrari *et al.*, 1992). It was hypothesized by Ferrari and coworkers that this binding site provided a lower resistance to structural deformation and thus lowered the high energetic cost of bending DNA with an already bent substrate. Nagaich *et al.* (1997) examined the role of p53 in DNA bending and deduced that p53 caused bending in a number of DNA response elements. EMSA assays indicated that increasingly bent DNA had a higher affinity interaction with p53. Whilst the examined response elements did not contain strict AT tracts, they did contain dinucleotide AT regions on either side of the response element. TATA-box binding protein (TBP) binding to its DNA target is largely mediated by the bent shape that its DNA target takes and this contributes to the affinity of the complex (Harteis & Schneider, 2014). The energetic cost of the shape of the DNA is lowered by the reduced stacking of the TA steps in the TATA sequence.

Realizing the pre-bent nature of many of these DNA targets is important in understanding the affinity and energetics of certain protein–DNA interactions. Protein recognition of AT-rich DNA can be complicated because of many interrelated effects. However, DNA bending is reportedly sensitive to environmental conditions such as temperature, salt and divalent cations (Haran & Mohanty, 2009). Given an initial hypothesis of unwinding, bending or fluctuational opening of dsDNA playing a role in the affinity of a protein–DNA complex, it may be possible to fine-tune environmental conditions to maximize such effects and so increase the affinity of protein–DNA interactions for structural studies. Furthermore, experimental data on DNA bending in response to sequence composition may also be used to train structure-prediction tools such as successors to *AlphaFold*3. Our polymer-model approach may also perhaps present an opportunity for the development of other multiphase modelling systems that can separately model both the DNA by itself and in complex with a protein. This could be technically possible via a contrast-matching small-angle neutron scattering experiment where an experimentalist could make the protein 'invisible' and only analyse the DNA in an unbound and also a protein-bound state.

4. Conclusion

In this work, we have shown that SAXS data can be analysed to determine the statistical properties of the conformation of short DNA sequences in solution. Although the SAXS structure factor averages over all spatial orientations of the molecules, for sequences which have some asymmetry, even when it is rather weak such as for GAGE6_3, the computed conformations can be oriented with respect to the sequence so that specific features detected by the analysis can be related to the DNA sequence.

Our analysis uses a polymer model which is simple enough to allow a very broad exploration of conformational space (up to 10^8 or 10^9 conformations with moderate computing facilities) but is nevertheless able to quantitatively describe the average persistence length and torsional rigidity of the DNA double helix. This is important to ensure that only conformations accessible to a DNA molecule are generated. In contrast to the protein chains in intrinsically disordered protein regions, which can be described by a random Gaussian chain (Martin *et al.*, 2021), the bending and torsional rigidities of the double helix introduce strong constraints on the conformational space, which are intrinsic to our polymer model. The model, which is restricted to a chain of rigid segments, does not impose any constraints tying bending to twist. Nevertheless, our analysis detects the increase of twist fluctuations which is expected for DNA undergoing strong bending due to the mechanical properties of the double helix. The ability of the analysis to detect a property of DNA which in itself is not built into the model strengthens the credibility of the method, as well as its capability to detect regions which are prone to bending in correlation with the sequence.

As expected, the results confirm that AT-rich regions are more flexible than GC-rich domains; however, our results have also shown that the link between the sequence and the preferred conformations of free DNA in solution is not trivially imposed by the local sequence. Using a series of sequences derived from one another by small local changes, we have exhibited unexpected properties which suggest that collective effects are also very important.

Our method relies on a fit of $P_{\rm exp}(r)$, *i.e.* analysis carried out in real space, rather than handling the data in reciprocal space. This introduces some difficulty because an accurate derivation of $P_{\rm exp}(r)$ from the experimental data is not straightforward, although it has been the object of numerous investigations. The determination of the regularization parameter α which enters into the calculation is delicate. However, our analysis has shown that when α is modified, the results on the properties of the sample are robust.

It turns out that the view in real space is essential to detect *local* effects which are spread out in Fourier space. This is not specific to SAXS and spans various domains of physics. A famous example is provided by nonlinear localized waves in

water or solid-state physics (Dauxois *et al.*, 2005; Dauxois & Peyrard, 2000). A numerical experiment on vibrational waves to study thermalization in solids by Enrico Fermi and coworkers raised a puzzling question that was unexplained for more than ten years. It remained a mystery because, being a wave problem, the results were displayed in the wavevector space (q space). It was only when some physicists looked in real space that they obtained a clue because they detected *localized* phenomena (later called solitons) which could easily solve the puzzle. The real and Fourier spaces bring complementary views, and although experiments and theory often lead naturally to the Fourier space, the real-space approach may sometimes be the most appropriate.

This work highlights the interest of SAXS studies for investigating protein binding sites on DNA because they are able to probe the conformations of DNA in solution and their dynamics, which is reflected in the statistics of the conformations provided by SAXS. Crystallographic and cryoEM studies, which are widely used to determine the structures of protein–DNA complexes, are very precise in their final structure, in which the conformational freedom of DNA is often highly constrained by being bound to a protein or vitrification and crystallization. However, when proteins scan DNA before binding, they often sample and test these free DNA conformations which can be observed by SAXS studies in solution. Therefore, SAXS, crystallographic and cryoEM studies are complementary. Having shown that the analysis of SAXS data is able to determine the properties of DNA in solution at a scale of a few base pairs enhances the interest in the combination of the two methods to investigate DNA–protein complexes.

Acknowledgements

Aspects of this research were undertaken on the SAXS/WAXS beamline at the Australian Synchrotron, Victoria, Australia, which is part of ANSTO, and we thank the beamline staff for their enthusiastic and professional support.

Funding information

This work was funded by the Australian Research Council (DP220103667, LE120100092 and LE140100096 to CSB) and the National Health and Medical Research Council of Australia (APP1147496 to CSB).

References

Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A. J., Bambrick, J., Bodenstein, S. W., Evans, D. A., Hung, C., O'Neill, M., Reiman, D., Tunyasuvunakool, K., Wu, Z., Žemgulytė, A., Arvaniti, E., Beattie, C., Bertolli, O., Bridgland, A., Cherepanov, A., Congreve, M., Cowen-Rivers, A. I., Cowie, A., Figurnov, M., Fuchs, F. B., Gladman, H., Jain, R., Khan, Y. A., Low, C. M. R., Perlin, K., Potapenko, A., Savy, P., Singh, S., Stecula, A., Thillaisundaram, A., Tong, C., Yakneen, S., Zhong, E. D., Zielinski, M., Žídek, A., Bapst, V., Kohli, P., Jaderberg, M., Hassabis, D. & Jumper, J. M. (2024). *Nature*, **630**, 493–500.

Anselmi, C., Bocchinfuso, G., De Santis, P. M. S., Savino, M. & Scipioni, A. (1999). *J. Mol. Biol.* **286**, 1293–1301.

Bernadó, P., Mylonas, E., Petoukhov, M., Blackledge, M. & Svergun, D. (2007). *J. Am. Chem. Soc.* **129**, 5656–5664.

Blanchet, C. & Svergun, D. (2013). *Annu. Rev. Phys. Chem.* **64**, 37–54.

Bryant, Z., Stone, M., Gore, J., Smith, S. B., Cozzarelli, N. R. & Bustamante, C. (2003). *Nature*, **424**, 338–341.

Chandrasekhar, S., Swope, T. P., Fadaei, F., Hollis, D. R., Bricker, R., Houser, D., Portman, J. J. & Schmidt, T. L. (2024). *bioRxiv*, 2024.02.14.579968.

Chirico, G., Collini, M., Tóth, K., Brun, N. & Langowski, J. (2001). *Eur. Biophys. J.* **29**, 597–606.

Chong, P. A., Vernon, R. M. & Forman-Kay, J. D. (2018). *J. Mol. Biol.* **430**, 4650–4665.

Crothers, D. (1988). *Proc. Natl Acad. Sci. USA*, **95**, 15613–15615.

Curuksu, J., Zacharias, M. L. R., Lavery, R. & Zakrzewska, K. (2009). *Nucleic Acids Res.* **37**, 3766–3773.

Dauxois, T. & Peyrard, M. (2000). *Physics of Solitons*, ch. 1 and ch. 8. Cambridge University Press.

Dauxois, T., Peyrard, M. & Ruffo, S. (2005). *Eur. J. Phys.* **26**, S3–S11.

Dickerson, R. & Drew, H. (1981). *J. Mol. Biol.* **149**, 761–786.

Ferrari, S., Harley, V. R., Pontiggia, A., Goodfellow, P. N., Lovell-Badge, R. & Bianchi, M. E. (1992). *EMBO J.* **11**, 4497–4506.

Franke, D., Jeffries, C. & Svergun, D. (2015). *Nat. Methods*, **12**, 419–422.

Gartenberg, M. R. & Crothers, D. M. (1988). *Nature*, **333**, 824–829.

Glatter, O. (1977). *J. Appl. Cryst.* **10**, 415–421.

Gorin, A. A., Zhurkin, V. & Olson, W. K. (1995). *J. Mol. Biol.* **247**, 34–48.

Grant, T. D., Luft, J. R., Carter, L. G., Matsui, T., Weiss, T. M., Martel, A. & Snell, E. H. (2015). *Acta Cryst.* D**71**, 45–56.

Guéron, M., Kochoyan, M. & Leroy, J.-L. (1987). *Nature*, **328**, 89–92.

Haran, T. E. & Mohanty, U. (2009). *Q. Rev. Biophys.* **42**, 41–81.

Harteis, S. & Schneider, S. (2014). *Int. J. Mol. Sci.* **15**, 12335–12363.

Heumann, H., Ricchetti, M. & Werel, W. (1988). *EMBO J.* **7**, 4379–4381.

Hizver, J., Rozenberg, H., Frolow, F., Rabinovich, D. & Shakked, Z. (2001). *Proc. Natl Acad. Sci. USA*, **98**, 8490–8495.

Hopkins, J. B., Gillilan, R. E. & Skou, S. (2017). *J. Appl. Cryst.* **50**, 1545–1553.

Kirby, N., Cowieson, N., Hawley, A. M., Mudie, S. T., McGillivray, D. J., Kusel, M., Samardzic-Boban, V. & Ryan, T. M. (2016). *Acta Cryst.* D**72**, 1254–1266.

Knott, G., Bond, C. & Fox, A. (2016). *Nucleic Acids Res.* **44**, 3989–4004.

Koning, H. J., Lai, J. Y., Marshall, A. C., Stroeher, E., Monahan, G., Pullakhandam, A., Knott, G. J., Ryan, T. M., Fox, A. H., Whitten, A., Lee, M. & Bond, C. S. (2025). *Nucleic Acids Res.* **53**, gkae1198.

Lankaš, F., Šponer, J. L. J., Langowski, J. & Cheatham, T. E. III (2003). *Biophys. J.* **85**, 2872–2883.

Lavery, R., Zakrzewska, K., Beveridge, D., Bishop, T. C., Case, D. A., Cheatham, T. III, Dixit, S., Jayaram, B., Lankas, F., Laughton, C., Maddocks, J. H., Michon, A., Osman, R., Orozco, M., Perez, A., Singh, T., Spackova, N. & Sponer, J. (2010). *Nucleic Acids Res.* **38**, 299–313.

Lee, M., Sadowska, A., Bekere, I., Ho, D., Gully, B. S., Lu, Y., Iyer, K. S., Trewhella, J., Fox, A. H. & Bond, C. S. (2015). *Nucleic Acids Res.* **43**, 3826–3840.

Lowary, P. & Widom, J. (1998). *J. Mol. Biol.* **276**, 19–42.

Lu, Y., Weers, B. & Stellwagen, N. (2002). *Biopolymers*, **61**, 261–275.

Manalastas-Cantos, K., Konarev, P. V., Hajizadeh, N. R., Kikhney, A. G., Petoukhov, M. V., Molodenskiy, D. S., Panjkovich, A., Mertens, H. D. T., Gruzinov, A., Borges, C., Jeffries, C. M., Svergun, D. I. & Franke, D. (2021). *J. Appl. Cryst.* **54**, 343–355.

Martin, E., Hopkins, J. & Mittag, T. (2021). *Methods Enzymol.* **646**, 185–222.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. & Teller, E. (1953). *J. Chem. Phys.* **21**, 1087–1092.

Nagaich, A. K., Appella, E. & Harrington, R. E. (1997). *J. Biol. Chem.* **272**, 14842–14849.

Nott, T. J., Craggs, T. D. & Baldwin, A. J. (2016). *Nat. Chem.* **8**, 569–575.
Olson, W., Gorin, A., Lu, X.-J., Hock, L. & Zhurkin, V. (1998). *Proc. Natl Acad. Sci. USA*, **95**, 11163–11168.
Panjkovich, A. & Svergun, D. (2018). *Bioinformatics*, **34**, 1944–1946.
Peterlin, A. (1953). *Nature*, **171**, 259–260.
Petoukhov, M. V., Franke, D., Shkumatov, A. V., Tria, G., Kikhney, A. G., Gajda, M., Gorba, C., Mertens, H. D. T., Konarev, P. V. & Svergun, D. I. (2012). *J. Appl. Cryst.* **45**, 342–350.
Rambo, R. & Tainer, J. (2013). *Nature*, **496**, 477–481.
Rohs, R., West, S. M., Sosinsky, A., Liu, P., Mann, R. S. & Honig, B. (2009). *Nature*, **461**, 1248–1253.
Ryan, T. M., Trewhella, J., Murphy, J. M., Keown, J. R., Casey, L., Pearce, F. G., Goldstone, D. C., Chen, K., Luo, Z., Kobe, B., McDevitt, C. A., Watkin, S. A., Hawley, A. M., Mudie, S. T., Samardzic Boban, V. & Kirby, N. (2018). *J. Appl. Cryst.* **51**, 97–111.
Saenger, W. (1984). *Principles of Nucleic Acid Structure.* New York: Springer-Verlag.
Schellman, J. (1974). *Biopolymers*, **13**, 217–226.
Schindler, T., González, A., Boopathi, R., Roda, M. M., Romero-Santacreu, L., Wildes, A., Porcar, L., Martel, A., Theodorakopoulos, N., Cuesta-López, S., Angelov, D., Unruh, T. & Peyrard, M. (2018). *Phys. Rev. E*, **98**, 042417.
Schneidman-Duhovny, D., Hammel, M., Tainer, J. & Sali, A. (2016). *Nucleic Acids Res.* **44**, W424–W429.
Sivia, D. (2011). *Elementary Scattering Theory. For X-ray and Neutron Users.* Oxford University Press.
Smith, S., Finzi, L. & Bustamante, C. (1992). *Science*, **258**, 1122–1126.
Song, X., Sun, Y. & Garen, A. (2005). *Proc. Natl Acad. Sci. USA*, **102**, 12189–12193.
Stefl, R., Wu, H., Ravindranathan, S., Sklenář, V. & Feigon, J. (2004). *Proc. Natl Acad. Sci. USA*, **101**, 1177–1182.
Svergun, D. I. (1992). *J. Appl. Cryst.* **25**, 495–503.
Svergun, D., Barberato, C. & Koch, M. H. J. (1995). *J. Appl. Cryst.* **28**, 768–773.
Theodorakopoulos, N. & Peyrard, M. (2012). *Phys. Rev. Lett.* **108**, 078104.
Trewhella, J., Duff, A. P., Durand, D., Gabel, F., Guss, J. M., Hendrickson, W. A., Hura, G. L., Jacques, D. A., Kirby, N. M., Kwan, A. H., Pérez, J., Pollack, L., Ryan, T. M., Sali, A., Schneidman-Duhovny, D., Schwede, T., Svergun, D. I., Sugiyama, M., Tainer, J. A., Vachette, P., Westbrook, J. & Whitten, A. E. (2017). *Acta Cryst.* D**73**, 710–728.
Tria, G., Mertens, H. D. T., Kachala, M. & Svergun, D. I. (2015). *IUCrJ*, **2**, 207–217.
Urban, R., Bodenburg, Y. & Wood, T. (2002). *Am. J. Physiol. Endocrinol. Metab.* **283**, E423–E427.
Wang, J., Sachpatzidis, A., Christian, T. D., Lomakin, I. B., Garen, A. & Konigsberg, W. H. (2022). *Biochemistry*, **61**, 1723–1734.
Widom, J. (1984). *Nature*, **309**, 312–313.